MARIANNE SCHAEFER FRANÇA

ANÁLISE ESTATÍSTICA MULTIVARIADA DOS DADOS DE MONITORAMENTO DE QUALIDADE DE ÁGUA DA BACIA DO ALTO IGUAÇU: UMA FERRAMENTA PARA A GESTÃO DE RECURSOS HÍDRICOS

CURITIBA
2009

MARIANNE SCHAEFER FRANÇA

ANÁLISE ESTATÍSTICA MULTIVARIADA DOS DADOS DE MONITORAMENTO DE QUALIDADE DE ÁGUA DA BACIA DO ALTO IGUAÇU: UMA FERRAMENTA PARA A GESTÃO DE RECURSOS HÍDRICOS

Dissertação apresentada ao Curso de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia.

Orientador: Cristovão V. S. Fernandes, Ph.D.

Co-orientador: Eloy Kaviski, Dr.

CURITIBA

2009

França, Marianne Schaefer
Análise estatística multivariada dos dados de monitoramento de qualidade de água da Bacia do Alto Iguaçu: uma ferramenta para a gestão de recursos hídricos / Marianne Schaefer França – Curitiba, 2009.
150 f. : il., tabs, grafs.

Orientador: Cristovão V. S. Fernandes
Co-Orientador: Eloy Kaviski
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Tecnologia, Curso de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental.

1. Água – Controle de qualidade. 2. Análise multivariada. 3. Bacias hidrográficas. IV. Recursos hídricos - Desenvolvimento. I. Fernandes, Cristovão V. S.. II. Kaviski, Eloy. III. Título. IV. Universidade Federal do Paraná.

CDD 551.48

TERMO DE APROVAÇÃO

MARIANNE SCHAEFER FRANÇA

"ANÁLISE ESTATÍSTICA MULTIVARIADA DOS DADOS DE MONITORAMENTO DE QUALIDADE DE ÁGUA DA BACIA DO ALTO IGUAÇU: UMA FERRAMENTA PARA A GESTÃO DE RECURSOS HÍDRICOS".

Dissertação aprovada como requisito parcial à obtenção do grau de Mestre, pelo Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental do Setor de Tecnologia da Universidade Federal do Paraná, pela comissão formada pelos professores:

PRESIDENTE:

Cristovão Vicente Scapulatempo Fernandes
Universidade Federal do Paraná
Orientador

MEMBROS:

Eloy Kaviski
Universidade Federal do Paraná
Co-orientador

Monica Ferreira do Amaral Porto
Universidade de São Paulo

Júlio Cesar Rodrigues de Azevedo
Universidade Tecnológica Federal do Paraná

Jair Mendes Marques
Universidade Federal do Paraná

Curitiba, 14 de maio de 2009

AGRADECIMENTOS

A Deus, por me reservar sempre boas surpresas e me cercar de pessoas muito especiais.

Ao professor Cristovão, por todo aprendizado, incentivo, paciência, dedicação, preocupações e pelas diversas oportunidades. Obrigada professor por dividir o fardo comigo :)

Ao meu noivo, familiares e amigos por tornar esta caminhada mais fácil, pelos momentos divertidos e de descontração!

À Heloise Knapik, pela amizade e companhia, pelos pães de queijo, pães de mel, pelas trufas e festas! Muito obrigada também pela paciência e ajuda nos momentos de dúvidas cruéis ;)

À Vanessa Gonçalves e à Clarissa Scuissiato, por se disponibilizarem a ajudar nas campanhas de monitoramento e no laboratório.

Ao professor Jair Mendes Marques, por suas aulas que muito me auxiliaram na elaboração deste trabalho e por sempre estar à disposição para sanar minhas dúvidas.

Ao professor Eloy Kaviski, por toda atenção e pelo tempo tomado para discussão deste trabalho.

Ao CNPq/CT-Hidro, pela bolsa concedida para a realização desta pesquisa.

SUMÁRIO

CAPÍTULO III - AVALIAÇÃO DA QUALIDADE DA ÁGUA EM BACIAS CRÍTICAS: ESTRATÉGIA PARA AVALIAÇÃO ESTATÍSTICA

CAPÍTULO IV - RESULTADOS

CAPÍTULO V – CONCLUSÕES E RECOMENDAÇÕES

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AA	Análise de Agrupamentos
ACP / PCA	Análise de Componentes Principais / Principal Component Analysis
AF / FA	Análise Fatorial / Factor Analysis
DBO_5	Demanda Bioquímica de Oxigênio
DQO	Demanda Química de Oxigênio
Cond	Condutividade
COT	Carbono Orgânico Total
CP	Componente Principal
F	Fator
Fósf	Fósforo
KMO	Medida de Adequacidade da Amostra de Kaiser-Meyer-Olkin
MV	Máxima Verossimilhança
N-A	Nitrogênio Amoniacal
N-Org	Nitrogênio Orgânico
NO_2^-	Nitrito
NO_3^-	Nitrato
OD	Oxigênio Dissolvido
pH	Potencial Hidrogeniônico
Q	Vazão
r	correlação
RMC	Região Metropolitana de Curitiba
Secchi	Profundidade do Disco de Secchi
SDT	Sólidos Dissolvidos Totais
SST	Sólidos Suspensos Totais
SSed	Sólidos Sedimentáveis
T	Temperatura da Água
Turb	Turbidez

LISTA DE SÍMBOLOS

μ	Média
ρ	Matriz de correlação
Σ	Matriz de covariância
α	Nível de significância
ν	Grau de liberdade

RESUMO

O principal objetivo deste trabalho foi realizar a análise multivariada dos dados de monitoramento de qualidade de água da bacia do Alto Iguaçu, utilizando-se das seguintes técnicas: Análise de Componentes Principais, Análise Fatorial e Análise de Agrupamentos. Adotaram-se duas estratégias de avaliação, a primeira refere-se à Análise Global da Bacia do Alto Iguaçu. Nesta análise as variáveis avaliadas foram 18 parâmetros de qualidade de água, incluindo a vazão. O objetivo foi identificar quais parâmetros seriam mais relevantes para caracterização do estado qualitativo do corpo hídrico. Para tanto, foram utilizadas as técnicas das Componentes Principais e Fatorial, empregando-se os softwares MATLAB e STATISTICA. Os parâmetros considerados mais significantes foram o Oxigênio Dissolvido (OD), o Nitrogênio Amoniacal, a Condutividade, o pH, os Sólidos Suspensos Totais, o Nitrogênio Orgânico e a Turbidez, os quais destacaram os aspectos de degradação da matéria orgânica e sua interação com a dinâmica de transporte de sólidos. A segunda estratégia adotada foi a Análise dos Pontos de Monitoramento da Bacia do Alto Iguaçu realizada através da Análise de Componentes Principais, com o objetivo de levantar quais pontos de amostragem seriam mais representativos para o monitoramento da bacia e a relação existente entre estes pontos. Foram selecionadas as duas primeiras componentes principais, que em conjunto explicaram cerca de 97% da variância da amostra. A CP1 agrupou os pontos P2 a P6, sendo estes considerados os mais relevantes, mostrando que o resultado tendeu para os pontos mais poluídos. Na CP2, foi possível observar o contraste entre o ponto P1 e os demais, mostrando justamente que este se diferencia dos demais por estar situado em uma área de manancial. Adicionalmente, realizou-se também a Análise de Agrupamentos das Coletas de Água. Foram obtidos dois grupos principais: o de coletas que refletiram melhor qualidade do corpo hídrico, formado principalmente por coletas realizadas no ponto P1, próximo a uma área de manancial da bacia; e, o outro formado por grande parte das outras coletas, as quais refletiram o estado de degradação do rio, evidenciando e confirmando que em sua totalidade, a qualidade da água da bacia apresenta-se inadequada.

Palavras-Chave: Qualidade da Água, Análise Multivariada, Gestão de Recursos Hídricos.

ABSTRACT

This work presents the strategies used to apply the concepts of multivariate analysis for water quality monitoring data of the Iguaçu River at the Metropolitan Area of Curitiba, considering three distinct techniques: Principal component Analysis (PCA), Factor Analysis (FA) and Cluster Analysis. To achieve the main goals, two distinct evaluation strategies were used. In the first one, called Global Analysis, 18 water quality parameters were considered as variables including water flows. The goal was to identify which parameters would better represent the water quality condition of a given water resource, based upon the use of the PCA and FA techniques developed through routines inside the MATLAB and STATISTICA softwares. The most significant water quality parameters are: Dissolved Oxygen (DO), Amoniacal Nitrogen ($N\text{-}NH_3$), Organic Nitrogen, Conductivity, pH and Total Supended Solids. This result highliths the impact of the organic content in the river and its interaction with the solid transport dynamic. The second strategy was based on use of Principal Component Analysis for the monitoring points of the Iguaçu River aiming to define the most representatives for monitoring purposes and its main relations. The 2 first principal components were chosen to explain 97% of sample variance. CP1 involves P2 to P6 as the most relevant, indicating the monitoring points at the most polluted areas. CP2 allowed evaluating the contrast between P1 and the others, revealing the influence of the watershed area. Additionally, the cluster analysis was used to evaluate the impact of the sampling process. Two main results were obtained: sampling reproducing the good water quality condition of the most upstream monitoring point (P1) in the water supply area. The other cluster indicates that the sampling process reproduce the water quality degradation of the Iguaçu River.

Key-words: Water quality, Multivariate Analysis, Water Resources Management.

CAPÍTULO I

1. INTRODUÇÃO

As bacias hidrográficas geralmente constituem áreas com disponibilidade de regiões férteis que contemplam a atividade agrícola e áreas que possibilitam o sustento de diversos usos como a irrigação, o abastecimento industrial e o doméstico. Adicionalmente, o rio desempenha um papel importante quanto ao transporte e à assimilação de efluentes domésticos e industriais, bem como aqueles resultantes do escoamento de áreas agrícolas, estradas e avenidas. No entanto, o uso abusivo do corpo hídrico acaba por comprometer a sua qualidade, o que afeta diretamente alguns usos a que a bacia se propõe.

No caso do comportamento hidrológico da bacia, este depende de fatores como a precipitação, contribuição de vazão de afluentes, escoamento superficial, clima, entre outros. Além disso, as variações destes fatores exercem influências sobre a vazão da bacia, que está diretamente ligada com a concentração dos poluentes na água do rio. Ou seja, a dinâmica de uma bacia apresenta-se de forma complexa, o que torna difícil compreender e prever o seu comportamento.

Deste modo, pesquisas de longo prazo e programas de monitoramento de qualidade da água são primordiais para um melhor entendimento sobre o comportamento do corpo hídrico. Segundo Tucci (2001), as informações hidrometeorológicas e de qualidade da água são indispensáveis para se promover um adequado aproveitamento dos recursos hídricos em bases sustentáveis. A falta de informações aumenta a incerteza nas decisões, acarretando resultados negativos no uso e aproveitamento dos recursos hídricos. De um modo geral, o custo associado à falta das informações é geralmente superior ao custo da obtenção do dado e de sua análise final em um projeto.

Além disso, de acordo com Brito *et al*. (2003), a implementação de políticas de gestão e monitoramento da qualidade das águas são ações prioritárias para auxiliar na definição de medidas de prevenção e conservação dos recursos hídricos, que visem à melhoria da qualidade da água e conseqüentemente aumento da disponibilidade.

No Brasil, a situação do monitoramento de qualidade da água é bastante deficitária. Segundo Porto (2003), um dos maiores déficits do país na área de qualidade da água está na aquisição e utilização da informação. Faltam redes de monitoramento de qualidade da água, a infra-estrutura laboratorial é insuficiente e há dificuldades na análise e divulgação destas

informações. É inegável que também há forte déficit de capacitação no setor. Finalmente, são poucos os grupos de pesquisa que trabalham com aspectos de qualidade da água dos corpos hídricos, certamente em menor número do que a nossa extensa rede hídrica demandaria.

Um levantamento divulgado pelo Ministério do Meio Ambiente no ano de 2002 indica que apenas São Paulo, Minas Gerais e Mato Grosso do Sul classificam-se em boa situação e, no outro extremo, Acre, Alagoas, Amazonas, Ceará, Maranhão, Pará, Paraíba, Piauí, Rio Grande do Norte, Rondônia, Roraima, Santa Catarina e Sergipe classificam-se em situação de monitoramento incipiente.

Complementarmente, o *site* do IBAMA na *internet* indica existirem 1.985 estações de monitoramento de qualidade da água no Brasil, sendo que destas, 1.241 continuam operando, isto é são estações ativas. Nos Estados Unidos, a título de exemplo, a Agência de Proteção Ambiental (EPA) tem registrado no seu site na Internet a existência de 134.858 estações de monitoramento de qualidade da água e permite que qualquer entidade que opere estações de monitoramento inclua seus dados no seu site (PORTO, 2003).

Em países desenvolvidos, contudo, onde geralmente há maiores investimentos em monitoramento e gestão da qualidade das águas, um problema aparente é o grande conjunto de dados de qualidade de água gerados e a dificuldade freqüente que existe em interpretá-los (DIXON & CHISWELL[1],1996, citado por VEGA *et al*., 1998). Deste modo, fica evidente que o problema quanto à compreensão do comportamento do corpo hídrico não se limita apenas à disponibilidade de estações de monitoramento e laboratórios qualificados. É necessário compreender o significado das variáveis de qualidade de água e as suas interações, bem como a resposta da bacia aos diversos processos que ocorrem na sua superfície.

Uma alternativa para compreensão do significado dos dados de qualidade de água é a análise estatística, conforme dispõe a própria Resolução CONAMA 357/05, Artigo 8º, Capítulo III: "§ 2º Os resultados do monitoramento deverão ser analisados estatisticamente e as incertezas de medição consideradas". No entanto, o emprego da estatística clássica - do ponto de vista prático - não seria muito esclarecedor, visto que para a avaliação da qualidade da água seria necessário estudar a relação entre muitas variáveis: os diversos parâmetros de qualidade de água.

Assim sendo, o emprego de técnicas multivariadas seria o mais indicado, visto que uma de suas atribuições é analisar grandes conjuntos de dados referentes a diversas variáveis.

[1] DIXON, W. & CHISWELL, B. **Review of aquatic monitoring program design**. Water Resources , nº 30, p. 1935-1948, 1996.

Além disso, através das análises multivariadas é possível simplificar a estrutura de variabilidade dos dados (MINGOTI, 2005), facilitando a interpretação dos mesmos.

Outra vantagem do emprego de técnicas de análise estatística multivariada, segundo Nonato *et al*. (2007), é a possível otimização da rede de amostragem proposta bem como da freqüência de amostragem e do número de parâmetros analisados, sem perda de informação, visto que programas de monitoramento são dispendiosos.

Assim, o enfoque do presente trabalho não se refere às necessidades básicas do país quanto ao monitoramento, que seriam a expansão da rede e do número de laboratórios capacitados entre outros, mas sim o destino e o significado dos dados, as possíveis relações existentes entre as variáveis e a possibilidade de reduzir o número de parâmetros monitorados, potencializando aqueles com maior contribuição para a qualidade da água, reduzindo custos, tempo gasto em campanhas de monitoramento e em laboratório.

1.1 JUSTIFICATIVA

O despertar do interesse a respeito de uma abordagem estatística mais detalhada dos dados de monitoramento da bacia do Alto Iguaçu decorreu dos resultados de KNAPIK (2006) acerca do cálculo do coeficiente de correlação de Pearson (r) para dados monitorados na mesma bacia. O coeficiente de Pearson é um indicador que descreve a interdependência linear entre duas variáveis x e y, e pode ser calculado pela seguinte equação, onde $\overline{x}$ e $\overline{y}$ são as médias das variáveis em estudo:

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}} \qquad (1.1)$$

No estudo de Knapik (2006) foram adotados intervalos que relacionavam os valores de r com o tipo de correlação (fraca, moderada, forte, etc.). Estes intervalos são tradicionalmente utilizados na literatura. No entanto, há fatores que podem afetar a intensidade do coeficiente de Pearson, como o tamanho da amostra, a existência de valores muito discrepantes, a restrição da amplitude de uma das variáveis ou de ambas, além dos erros de medição, o que sugere que adotar estes tipos de intervalos nem sempre é o mais adequado, visto que isto pode levar a interpretações mais subjetivas dos resultados.

No caso dos parâmetros de qualidade de água, que em conjunto auxiliam a compreender o estado e o comportamento de um sistema hídrico - o que é algo complexo justamente por envolver tantas variáveis, não é fácil nem cauteloso afirmar que as relações entre parâmetros são fracas ou fortes, visto que medem objetos diferentes em escalas diferentes.

Deste modo, o que se visa evidenciar é que se basear em resultados da estatística clássica nem sempre é suficiente, além de trazer muitas vezes considerações subjetivas que exigem complementações para que se possa chegar a resultados mais conclusivos.

Assim, surgiu o interesse em se aplicar a análise estatística multivariada nos dados de qualidade de água monitorados na bacia do Alto Iguaçu na Região Metropolitana de Curitiba. A análise estatística multivariada é na verdade um conjunto de distintas técnicas que revelam mais informações do que a estatística clássica, além de permitir utilizar os métodos para diversas variáveis simultaneamente.

Além disso, é raro encontrar na literatura procedimentos detalhados para aplicação deste tipo de análise em dados de monitoramento de qualidade de água. E, como justificativa final, tem-se que este será o primeiro estudo realizado na Região Metropolitana de Curitiba quanto à aplicação de técnicas multivariadas de dados de monitoramento de qualidade de água da bacia do Rio Iguaçu, que abrange grande parte desta Região, com enfoque na definição de estratégias para a gestão de recursos hídricos.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Aprofundar os conhecimentos relacionados à análise multivariada aplicada à gestão da qualidade das águas, destacando seus benefícios e limitações, fornecendo assim, subsídios técnicos consistentes que sirvam de orientação para os comitês de bacias hidrográficas e órgãos gestores de recursos hídricos instituírem seus planos de bacias de uma forma realista e sustentável, realizando para tanto a análise multivariada dos dados de monitoramento de qualidade de água da bacia do Alto Iguaçu.

1.2.2 Objetivos Específicos

Os objetivos específicos deste estudo são apresentados na sequência:

1) Realizar a análise multivariada dos dados de qualidade de água monitorados na Bacia do Alto Iguaçu considerando as seguintes técnicas estatísticas: Análise de Componentes Principais (ACP), Análise Fatorial (AF) e Análise de Agrupamentos (AA). Para tanto, serão utilizadas rotinas estatísticas nos *softwares* MATLAB e STATISTICA.
2) A partir das análises de componentes principais e fatorial, indicar um conjunto representativo de parâmetros de qualidade de água que possam eventualmente mostrar a melhor estratégia de monitoramento. Ou seja, no caso de não haver disponibilidade para realizar o monitoramento de todos os parâmetros de qualidade normalmente monitorados, é interessante monitorar ao menos aqueles considerados mais relevantes.
3) Realizar a análise de componentes principais, considerando como variáveis os pontos de monitoramento, buscando identificar as relações existentes entre eles, bem como os pontos de monitoramento mais relevantes para a avaliação da qualidade da água.
4) A partir da análise de agrupamentos, reunir em grupos as coletas de amostras de água do rio, visando encontrar aquelas que refletiram melhor e pior qualidade do corpo hídrico.

1.3 MÉTODO

O desenvolvimento deste trabalho seguiu três etapas: (i) realização de coletas de amostras de água nos 7 pontos de monitoramento da bacia do Alto Iguaçu visando obter um maior conjunto de dados, a partir da complementação do conjunto de amostras de água obtido durante a realização do Projeto Bacias Críticas (PORTO, 2007); (ii) realização das análises das amostras de água em laboratório de acordo com o *Standard Methods* (APHA, 1998) e (iii) aplicação dos métodos multivariados considerando o conjunto de dados obtido. Para tanto, foram utilizadas três técnicas multivariadas: Análise de Componentes Principais (ACP), Análise Fatorial (AF) e Análise de Agrupamentos (AA). Para a aplicação da ACP, foram utilizadas rotinas programadas no *software* MATLAB versão 5.3. Na AF, foram realizados alguns testes previamente no MATLAB para verificação das condições de uso da análise. A AF propriamente dita foi realizada no *software* STATISTICA versão 6.0. Para a análise de agrupamentos também utilizou-se o *software* STATISTICA, além de algumas rotinas auxiliares programadas no MATLAB.

1.4 ESTRUTURA DA DISSERTAÇÃO

A presente dissertação está estruturada em cinco capítulos, sendo estes: Capítulo I – Introdução, Capítulo II – Aspectos Conceituais da Análise Multivariada, Capítulo III – Abordagem Metodológica para Aplicação da Análise Multivariada para a Gestão de Recursos Hídricos, Capítulo IV – Resultados e Análises, Capítulo V – Conclusões e Recomendações.

O Capítulo I aborda a análise estatística multivariada no contexto da gestão de recursos hídricos bem como sua importância como instrumento de suporte à decisão na gestão de qualidade da águas. Contém a justificativa e os objetivos do trabalho, e, apresenta o método adotado para a realização deste estudo.

O Capítulo II refere-se aos aspectos conceituais da análise multivariada, abordando as técnicas estatísticas a serem utilizadas: Análise de Componentes Principais, Análise Fatorial e a Análise de Agrupamentos. Complementarmente apresenta experiências de outros autores, visando elucidar a aplicabilidade da análise multivariada em dados de monitoramento de qualidade de água.

No Capítulo III, apresentam-se a bacia do Alto Iguaçu e os pontos de monitoramento localizados em sua extensão, bem como os parâmetros de qualidade de água utilizados para avaliação qualitativa do corpo hídrico. Discute-se também a aplicação propriamente dita dos métodos propostos no Capítulo II e as estratégias de avaliação dos dados monitorados.

O Capítulo IV exibe os resultados obtidos e suas respectivas análises, de acordo com os objetivos propostos.

O Capítulo V apresenta as conclusões referentes aos resultados obtidos e algumas recomendações.

Adicionalmente, constam materiais referentes às funções programadas utilizadas no software MATLAB versão 5.3, dados de monitoramento de qualidade e quantidade de água referentes a cada um dos pontos de monitoramento, algumas fotos das campanhas de monitoramento e resultados complementares.

CAPÍTULO II

2. ASPECTOS CONCEITUAIS DE ANÁLISE ESTATÍSTICA MULTIVARIADA NA AVALIAÇÃO DA QUALIDADE DA ÁGUA

A dinâmica de uma bacia hidrográfica apresenta-se de forma complexa, tornando difícil compreender e prever o seu comportamento. Assim sendo, pesquisas de longo prazo e programas de monitoramento de qualidade e quantidade de água fazem-se necessários para um maior entendimento acerca dos aspectos quali-quantitativos de um corpo hídrico. O resultado de programas de monitoramento mais longos é um grande conjunto de dados de diversos parâmetros de qualidade de água. E, por se tratar de um conjunto formado por diversas variáveis, medidas em diferentes escalas e unidades, sua interpretação não é trivial.

Deste modo, o emprego da estatística clássica não seria o mais indicado para avaliar este problema, mas sim, o uso de técnicas estatísticas multivariadas, capazes de analisar dados de diversas variáveis e locais simultaneamente.

Vega *et al*. (1998), por exemplo, realizaram a análise de seu conjunto de dados formado por 22 variáveis físicas e químicas com valores referentes a 3 pontos de monitoramento - o que resultou em um total de 30 amostras - através do uso de *box plots*, e, das técnicas multivariadas da ANOVA (Análise de Variância), da ACP (Análise de Componentes Principais) e da Análise de Agrupamentos. Foram identificados três grupos principais de parâmetros de qualidade de água, os quais os autores designaram por conteúdo mineral, poluição antropogênica e temperatura da água. Além disso, fontes temporais (sazonalidade e clima) e espaciais (poluição de fonte antropogênica) que afetam as características do corpo hídrico foram diferenciadas e atribuídas às fontes de poluição. Segundo os autores, a aplicação das análises multivariadas resultou em uma importante classificação das amostras de água do rio baseada em critérios sazonais e espaciais. Vega *et al*. (1998) também demonstraram que quando aplicados os testes de normalidade para cada uma das estações individualmente, estes validaram as distribuições normais para a maioria das variáveis, o que indicou a existência de diferenças na composição da água entre as estações.

No trabalho de Ouyang (2005), o autor optou por avaliar 22 estações de monitoramento por meio da Análise de Componentes Principais e da Análise Fatorial contando com dados de 42 parâmetros de qualidade de água, utilizando para tanto a mediana dos dados.

Ou seja, considerou como variáveis as estações de monitoramento e não os parâmetros de qualidade de água como ocorre mais freqüentemente. Deste modo, descobriram-se quais eram as estações de monitoramento mais representativas e quais poderiam eventualmente não ser mais monitoradas.

Bengraïne & Marhaba (2003) apresentaram diversas estratégias para avaliação de sua base de dados formada por 19 parâmetros de qualidade de água - além da vazão – utilizando-se da Análise Fatorial. O objetivo era monitorar alterações espaciais e temporais na qualidade de água do rio Passaic, que conta com 12 estações de monitoramento em New Jersey. Os dados foram avaliados inclusive por estação do ano. Na conclusão, os autores ressaltaram a importância do monitoramento ambiental associado ao uso de técnicas multivariadas para melhor compreensão de um sistema de água complexo.

Deste modo, é possível perceber a disseminação do uso de técnicas estatísticas multivariadas na análise de dados de monitoramento de qualidade de água, com o objetivo de se conhecer os parâmetros de qualidade de água mais representativos e obter um maior entendimento sobre a dinâmica de um corpo hídrico.

Contudo, ainda são escassas as referências bibliográficas que apresentem a utilização de técnicas estatísticas multivariadas no contexto da gestão dos recursos hídricos de modo detalhado. Assim, nesta pesquisa, será dada ênfase a esta versão conceitual.

2.1 ANÁLISE MULTIVARIADA

A Análise Multivariada pode ser definida como um conjunto de métodos estatísticos capazes de analisar medidas de *n* variáveis simultaneamente, sendo extremamente útil a pesquisadores que buscam compreender grandes e complexos conjuntos de dados.

Em linhas gerais, os métodos de estatística multivariada são utilizados com o propósito de simplificar ou facilitar a interpretação do fenômeno em estudo através da construção de índices ou variáveis alternativas que sintetizem a informação original dos dados; construir grupos de elementos amostrais que apresentem similaridade entre si, possibilitando a segmentação do conjunto de dados original; investigar as relações de dependência entre as variáveis respostas associadas ao fenômeno e outros fatores (variáveis explicativas), muitas vezes, com objetivos de predição; e, comparar populações ou validar suposições de testes de hipóteses (MINGOTI, 2005).

De acordo com Hair Jr *et al*. (1987), o caráter multivariado consiste nas múltiplas variáveis estatísticas (combinações múltiplas de variáveis) e não apenas no número de variáveis e observações. Assim, para uma amostra ser considerada realmente multivariada, todas as variáveis devem ser variáveis aleatórias que se inter-relacionam de tal modo que seus diferentes efeitos não podem ser interpretados separadamente.

Para Hardyck & Petrinovich[2] (1976, citado por Hair Jr *et al.*, 1987), os métodos de análise multivariada irão predominar no futuro e resultarão em drásticas mudanças no modo de pensar dos pesquisadores e no modo em que eles planejam suas pesquisas.

Segundo Mardia, Kent e Bibby (1979), em geral, se há *n* observações, o_1, ..., o_n e *p* variáveis, x_1, ..., x_p, os dados contém "*np*" peças de informação. Isto pode ser convenientemente representado utilizando-se uma matriz de dados (n x p), onde cada linha corresponde às observações e cada coluna corresponde a uma variável. Geralmente a matriz de dados pode ser escrita do seguinte modo:

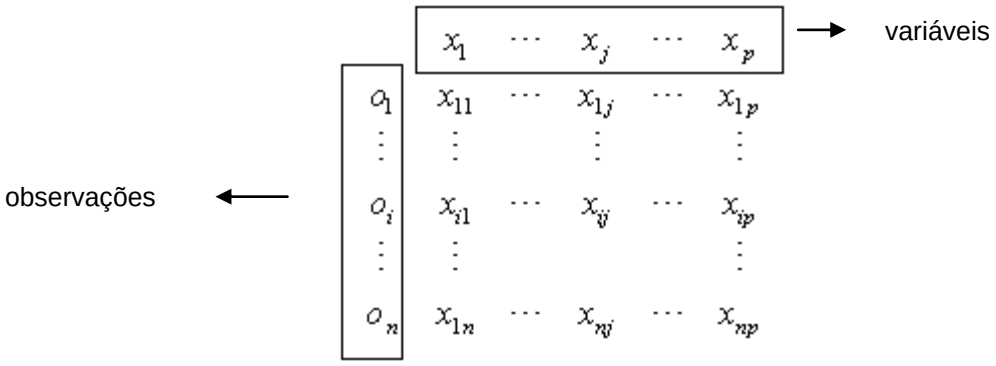


FIGURA 2.1 – Exemplo de matriz de dados
Fonte: Adaptado de MARDIA, KENT e BIBBY(1979)

Assim, a matriz de dados pode ser denotada por **X** (*n* x *p*), sendo representada como:

$$\mathbf{X}=\begin{bmatrix}\underline{x}_1' \\ \vdots \\ \underline{x}_n'\end{bmatrix}=\begin{bmatrix}x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & & x_{ij} & & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np}\end{bmatrix} \qquad (2.1)$$

Conforme Hair Jr. *et al.*[3] (2005, citado por Marques, 2006), na análise multivariada, a variável estatística pode ser definida como uma combinação linear de variáveis com pesos

[2] HARDYCK, C.D. & PETRINOVICH, L.F. **Introduction to Statistics for the Behavioral Sciences**. 2ª ed. Philadelphia: Saunders. 1976.

implicitamente determinados. Uma variável estatística de *n* variáveis ponderadas pode ser enunciada matematicamente como:

$$\text{Valor da variável estatística} = w_1X_1 + w_2X_2 + w_3X_3 + ... + w_nX_n \quad (2.2)$$

onde X_j é a variável observada e w_j, com j= 1, ..., n, é o peso determinado pela técnica multivariada. Tem-se como resultado um único valor que representa a combinação do conjunto inteiro de variáveis que melhor atinge o objetivo da análise multivariada específica.

Algumas técnicas de análise multivariada são: análise discriminante, análise de correlação canônica, regressão logística, análise de agrupamentos (ou *cluster*), análise multivariada da variância (MANOVA), análise fatorial e análise de componentes principais. Contudo, na presente dissertação serão abordadas somente as técnicas de análise de agrupamentos, de componentes principais e fatorial, em razão do tipo de resultado que estas técnicas fornecem e por se notar a preferência de muitos autores por estes tipos de análises em estudos semelhantes aos realizados no âmbito desta pesquisa (VEGA *et al.*, 1998; BENGRAÏNE & MARHABA, 2003; SHRESTA & KAZAMA, 2006; OUYANG, 2005; WUNDERLIN *et al.*, 2001; YEUNG, 1998).

2.1.1 Pré-requisitos para aplicação da análise multivariada

Para que o uso das técnicas multivariadas seja realizado de modo consistente há que se atentar a alguns pré-requisitos, os quais garantem a confiabilidade nos resultados obtidos.

Há, por exemplo, a questão relacionada ao número de variáveis (p) ser inferior ao número de observações (n), para qual ainda não existe uma resposta consensual. Ouyang (2005) afirmou em seu trabalho que se p > n, as soluções poderiam se tornar instáveis quando estimadas as matrizes de covariância e correlação na Análise de Componentes Principais ou na Análise Fatorial. Em contrapartida, apresentou também que outros estudos demonstraram que a ACP poderia ser aplicada a qualquer tipo de matriz e que estas discrepâncias poderiam ser devidas às diferentes soluções dos algoritmos utilizados nestes estudos. Já para Grossman[4] *et al*. (1991, citado por Yu *et al*., 1998), uma regra prática a ser adotada é a razão 3:1 (n:p) para que se obtenha uma solução estável na ACP.

---

[3] HAIR JR., J. F. et al. **Análise Multivariada de Dados**. 5 ed. Tradução: Adonai Schlup Sant'nna e Anselmo Chaves Neto. Porto Alegre: Bookman, 2005. Tradução de: Multivariate Analysis.

[4] GROSSMAN, G.D., NICKERSON, D.M. & FREEMAN, D.M. **Principal component analysis of assemblage structure data: untility of tests based oneigenvalues**. Ecology, 72, p. 341-347, 1991.

Neste trabalho, foi adotada a condição “n > p”, considerando que sob esta condição, o número de dados disponíveis (graus de liberdade) é maior, contudo a razão entre *n* e *p* (3:1) de Grossman não foi levada em consideração.

Outro questionamento se faz acerca dos testes que avaliam o grau de confiabilidade probabilística da análise fatorial em relação a diferentes bases de dados. Neste caso, citam-se o teste de esfericidade de Bartlett e a medida de adequacidade da amostra de Kaiser-Meyer-Olkin ou KMO (MINGOTI, 2005), que analisam se a estrutura de dados condiz com a análise fatorial e gerará então resultados mais confiáveis.

Além disso, para a aplicação de alguns testes e métodos, exige-se que os dados avaliados apresentem distribuição normal. O item a seguir trata deste assunto.

2.1.2 Distribuição Normal Multivariada

Segundo Marques (2006), a generalização da distribuição normal univariada para várias dimensões tem um papel fundamental na análise multivariada, pois grande parte das técnicas multivariadas aplicadas leva em consideração o fato de a amostra possuir distribuição normal multivariada. O método da verossimilhança para estimar os fatores da análise fatorial, por exemplo, exige que seja verificada a normalidade multivariada dos dados observados.

A densidade normal multivariada é uma generalização da densidade normal univariada *p* ≥ 2 dimensões. Denota-se, por conveniência, a função densidade de probabilidade da distribuição normal, com média μ e variância σ², por $X \sim N(\mu, \sigma^2)$. A distribuição normal univariada, com média μ e variância σ², tem função densidade de probabilidade dada por:

$$f_x(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in R,\ \mu \in R \text{ e } \sigma \in R^+ \tag{2.3}$$

A função densidade de probabilidade conjunta da normal com p variáveis independentes normais $X_1, X_2, ..., X_p$ tem a forma:

$$f(x_1, x_2,\ldots, x_p) = \frac{1}{(2\pi)^{\frac{p}{2}}\sigma_1\sigma_2\ldots\sigma_3}\exp\left[-\frac{1}{2}\sum_{i=1}^{p}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \tag{2.4}$$

Se $\underline{\mathbf{x}}' = \left[x_1, x_2, ..., x_p\right]$, $\underline{\boldsymbol{\mu}}' = \left[\mu_1, \mu_2, ..., \mu_p\right]$ e $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$, onde Σ é a matriz de covariância e $\sigma_{11} = \sigma_1^2$, $\sigma_{22} = \sigma_2^2$ e $\sigma_{pp} = \sigma_p^2$, pode-se escrever a densidade conjunta como:

$$f(\underline{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})\right] \tag{2.5}$$

onde: $-\infty < x_i < \infty$, i = 1, 2, ..., p.

Assumindo que **Σ (p x p)** é qualquer matriz simétrica positiva definida (2.6), obtém-se a função densidade geral da normal multivariada descrita em (2.5). Denota-se a função densidade normal p-dimensional por $\underline{\mathbf{X}} \sim \mathbf{N}_p(\underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{22} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} \tag{2.6}$$

Realizado o desenvolvimento da função densidade da normal multivariada $\underline{\mathbf{X}} \sim \mathbf{N}_p(\underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ com $\left|\boldsymbol{\Sigma}\right| > 0$, prova-se um resultado importante demonstrado em Johnson e Wichern (1998, p. 162-164) que é

$$\left(\underline{\mathbf{x}}\text{-}\underline{\boldsymbol{\mu}}\right)' \Sigma^{-1} \left(\underline{\mathbf{x}}\text{-}\underline{\boldsymbol{\mu}}\right) \sim \chi_p^2(\alpha) \tag{2.7}$$

com probabilidade 1-α, que pode-se denotar por

$$P\left[\left(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}\right)' \boldsymbol{\Sigma}^{-1} \left(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}\right) \leq \chi_p^2(\alpha)\right] = 1 - \alpha \tag{2.8}$$

onde χ_p^2 é obtido na tabela de distribuição de qui-quadrado, com p graus de liberdade.

Algumas propriedades da distribuição normal são fundamentais para o entendimento de modelos e métodos estatísticos. Com essas propriedades torna-se possível manipular as

distribuições normais facilmente o que a torna popular (MARQUES, 2006). As seguintes propriedades levam em consideração que **X** possui uma distribuição normal:

1) Combinações lineares das componentes de $\underline{\mathbf{X}} \sim N_p\left(\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$ são normalmente distribuídas.

2) Todos os subconjuntos das componentes de $\underline{\mathbf{X}} \sim N_p\left(\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$ tem uma distribuição normal (multivariada).

3) Covariâncias nulas implicam que as componentes correspondentes são independentemente distribuídas.

4) As distribuições condicionais das componentes de $\underline{\mathbf{X}} \sim N_p\left(\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$ são normais (multivariadas).

2.1.2.1 Avaliação da normalidade bivariada

Considerando $\underline{\mathbf{X}} \sim N_2\left(\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$ e substituindo no resultado (2.8) tem-se:

$$P\left[\left(\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}\right)' \Sigma^{-1}\left(\underline{\mathbf{x}}-\underline{\boldsymbol{\mu}}\right) \le \chi_2^2\left(0.5\right)\right] = 1-\alpha = 1-0.5 = 0.5 \qquad (2.9)$$

Com isso espera-se que 50% das observações amostrais situem-se dentro do contorno da elipse dada por

$$(\underline{\mathbf{x}}-\underline{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\underline{\mathbf{x}}-\underline{\bar{\mathbf{x}}}) \le \chi_2^2\left(0.5\right) \qquad (2.10)$$

onde se substitui $\underline{\boldsymbol{\mu}}$ pelo seu estimador $\underline{\bar{\mathbf{x}}}$ e Σ^{-1} pelo seu estimador $\mathbf{S}^{-1}$, caso contrário a hipótese de normalidade é suspeita.

2.1.2.2 Avaliação da normalidade de uma distribuição com p ≥ 2

De acordo com Johnson e Wichern (1998), um método mais formal para avaliar a normalidade de uma função com p ≥ 2 é baseado no quadrado da distância generalizada, dado por

$$d_j^2 = (\underline{\mathbf{x}}_j-\underline{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\underline{\mathbf{x}}_j-\underline{\bar{\mathbf{x}}}), \quad \text{com } j = 1, 2, ..., n \qquad (2.11)$$

onde $\underline{\mathbf{x}}_1$, $\underline{\mathbf{x}}_2$, ..., $\underline{\mathbf{x}}_n$ são as n observações amostrais.

Quando a população de onde a amostra foi retirada é normal multivariada e ambos “n” e “n–p” são maiores que 25, cada uma das distâncias $d_1^2, d_2^2, ..., d_n^2$ comportam-se como uma variável aleatória tipo qui-quadrado (χ^2) (MARQUES, 2003).

O método para avaliação da normalidade multivariada, neste caso, consiste nos seguintes passos:

1) Ordenar os quadrados das distâncias de forma crescente como $d_{(1)}^2 \leq d_{(2)}^2 \leq ... \leq d_{(n)}^2$.

2) Plotar os pares $\left[d_{(j)}^2, \chi_p^2 \left(\frac{j - \frac{1}{2}}{n} \right) \right]$.

Se o gráfico obtido resultar em uma linha reta aproximada, assume-se a normalidade, caso contrário rejeita-se a normalidade.

2.2 ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais é utilizada para a investigação das relações existentes em um conjunto de “p” variáveis, em geral, correlacionadas, transformando-o em um novo conjunto de variáveis não correlacionadas entre si denominadas componentes principais (CPs), onde estas são combinações lineares das “p” variáveis originais correlacionadas X_1, X_2, ..., X_p e possuem propriedades especiais em termos de variância.

Inicialmente o que se obtém da transformação das “p” variáveis originais correlacionadas são “p” componentes principais. No entanto, mesmo que sejam necessárias as “p” componentes principais para reproduzir a variabilidade total do sistema, a maior parte desta variabilidade pode ser explicada por um número menor “k” de componentes principais (k < p). Assim, como as “k” componentes principais explicam praticamente a mesma quantidade de informação que as “p” variáveis originais, podem-se substituir as “p” variáveis originais pelas “k” componentes principais, reduzindo-se o número de variáveis do problema em questão, perdendo no processo a menor quantidade de informação possível.

Geometricamente, as combinações lineares das variáveis originais representam a seleção de um novo sistema de coordenada obtido pela rotação do sistema original com coordenadas X_1, X_2, ..., X_p. Os novos eixos Y_1, Y_2, ..., Y_p representam a direção com

variabilidade máxima e permite uma interpretação mais simples da estrutura da matriz de covariância (JOHNSON & WICHERN, 1998). A título de exemplo, verifica-se na Figura 2.2 como fica rotação para o caso bivariado (p=2).

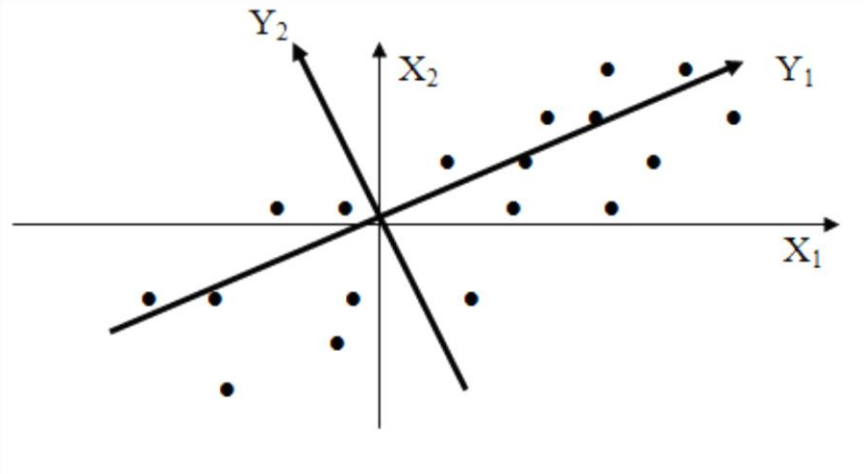


FIGURA 2.2 – Rotação para o caso bivariado
Fonte: MARQUES, 2003

De modo geral, os principais objetivos da análise de componentes principais são reduzir o número de variáveis, melhorar a interpretação e analisar quais variáveis ou conjuntos de variáveis explicam a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre elas. Além disso, a ACP é capaz de revelar informações que a princípio não se poderiam suspeitar. No entanto, segundo Johnson & Wichern (1998), esta análise funciona mais como um meio para o fim do que propriamente um fim, sendo muito útil como método auxiliar em Regressão, Análise Fatorial e Análise de Agrupamentos.

A obtenção das componentes principais depende somente da matriz de covariância "Σ" ou da matriz de correlação "ρ" de X_1, X_2, ..., X_p, não dependendo da suposição de normalidade (JOHNSON & WICHERN, 1998). Assim, o que ocorre quando a distribuição de probabilidades do vetor aleatório em estudo é normal p-variada, é que as componentes principais, além de serem não correlacionadas e independentes, têm distribuição normal.

A ACP é realizada a partir da matriz de correlação quando as unidades e escalas de mensuração são diferentes e no caso de uma variável apresentar variância muito maior do que as das outras. Outro modo equivalente para solucionar este problema é, por exemplo, primeiramente padronizar ou normalizar os dados (média = 0, variância = 1) e então realizar a análise a partir da matriz de covariância.

2.2.1 Componentes principais populacionais

Seja $\underline{\mathbf{X}}' = [X_1, X_2, ..., X_p]$ um vetor aleatório p-dimensional com vetor de médias $\underline{\mu}$ e matriz de covariância $\mathbf{\Sigma}$ com autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Considere as combinações lineares:

$$\begin{aligned} Y_1 &= \underline{\mathbf{a}}_1'\underline{\mathbf{X}} = a_{11}X_1 + a_{21}X_2 + ... + a_{p1}X_p \\ Y_2 &= \underline{\mathbf{a}}_2'\underline{\mathbf{X}} = a_{12}X_1 + a_{22}X_2 + ... + a_{p2}X_p \\ &\cdots\cdots\cdots\cdots \\ Y_j &= \underline{\mathbf{a}}_j'\underline{\mathbf{X}} = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p \\ &\cdots\cdots\cdots\cdots \\ Y_p &= \underline{\mathbf{a}}_p'\underline{\mathbf{X}} = a_{1p}X_1 + a_{2p}X_2 + ... + a_{pp}X_p \end{aligned} \qquad (2.12)$$

onde a_{ij}, com i = 1, 2, ..., p e j= 1, 2, ..., p.

Ou em notação vetorial

$$\underline{\mathbf{Y}} = \mathbf{A}'\underline{\mathbf{X}} \qquad (2.13)$$

onde

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} \qquad \text{e} \qquad A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \ldots & a_{pp} \end{bmatrix}$$

com

$$E(Y_i) = E(\underline{\mathbf{a}}_i'\underline{\mathbf{X}}) = \underline{\mathbf{a}}_i'E(\underline{\mathbf{X}}) = \underline{\mathbf{a}}_i'\underline{\mu} \qquad i = 1, 2, ..., p \qquad (2.14)$$

$$Var(Y_i) = V(\underline{\mathbf{a}}_i'\underline{\mathbf{X}}) = \mathbf{a}_i'V(\underline{\mathbf{X}})\underline{\mathbf{a}}_i = \mathbf{a}_i'\Sigma\mathbf{a}_i \qquad i = 1, 2, ..., p \qquad (2.15)$$

$$Cov(Y_i, Y_k) = V(\underline{\mathbf{a}}_i'\underline{\mathbf{X}}, \underline{\mathbf{a}}_k'\underline{\mathbf{X}}) = \underline{\mathbf{a}}_i'\Sigma\underline{\mathbf{a}}_k \qquad i \neq k = 1,2, ..., p \qquad (2.16)$$

onde $E(Y_i)$ é o valor esperado de Y_i, $Var(Y_i)$ é a variância de Y_i e $Cov(Y_i, Y_k)$ é a covariância entre Y_i e Y_k.

As componentes principais são as combinações lineares não correlacionadas Y_1, Y_2, ..., Y_p representadas em (2.12) e são derivadas em ordem decrescente de importância, ou seja, a primeira componente principal (Y_1) será responsável pela maior variância contida em todas as CPs e a última CP (Y_p), conseqüentemente, será responsável pela menor variância restante. As variâncias de cada componente principal são na verdade os autovalores (λ) da matriz de covariância - ou correlação dependendo do caso – sendo ordenadas do maior número para o menor.

Os coeficientes a_{ij}, i = 1, ..., p são denominados pesos ou carregamentos (*loadings*) das variáveis ou "fatores" e correspondem aos autovetores da matriz de covariância ou correlação dependendo do caso. Assim, quanto maior for o peso, maior será a importância da respectiva variável original (X_1, X_2, ..., X_p) na determinação da componente principal. Sinais positivos ou negativos indicam se a relação entre as variáveis originais e componentes principais é diretamente ou inversamente proporcional, respectivamente.

Pode-se definir que:

- A primeira componente principal é a combinação linear $\underline{\mathbf{a}}_1\underline{\mathbf{X}}$ que maximiza Var($\underline{\mathbf{a}}'_1\underline{\mathbf{X}}$) sujeito à condição $\underline{\mathbf{a}}'_1\underline{\mathbf{a}}_1 = 1$.
- A segunda componente principal é a combinação linear $\underline{\mathbf{a}}_2\underline{\mathbf{X}}$ que maximiza Var($\underline{\mathbf{a}}'_2\underline{\mathbf{X}}$) sujeito às condições $\underline{\mathbf{a}}'_2\underline{\mathbf{a}}_2 = 1$ e Cov($\underline{\mathbf{a}}'_1\underline{\mathbf{X}}$,$\underline{\mathbf{a}}'_2\underline{\mathbf{X}}$)= 0.
- A i-ésima componente principal é a combinação linear $\underline{\mathbf{a}}_i\underline{\mathbf{X}}$ que maximiza Var($\underline{\mathbf{a}}'_i\underline{\mathbf{X}}$) sujeito às condições $\underline{\mathbf{a}}'_i\underline{\mathbf{a}}_i = 1$ e Cov($\underline{\mathbf{a}}'_i\underline{\mathbf{X}}$,$\underline{\mathbf{a}}'_k\underline{\mathbf{X}}$)= 0 para k < i.

Para as componentes principais populacionais demonstram-se os seguintes resultados:

a) Seja o vetor aleatório $\underline{\mathbf{X}}' = [X_1, X_2, ..., X_p]$ associado a uma matriz de covariância $\mathbf{\Sigma}$ e pares de autovalores e autovetores (λ_1,$\underline{\mathbf{e}}_1$), (λ_2,$\underline{\mathbf{e}}_2$),..., (λ_p,$\underline{\mathbf{e}}_p$) onde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. Então a i-ésima componente principal é dada por:

$$Y_i = \underline{\mathbf{e}}'_i\underline{\mathbf{X}} = e_{1i}X_1 + e_{2i}X_2 + ... + e_{pi}X_p, \qquad i = 1, 2, ..., p \qquad (2.17)$$

com

$$Var(Y_i) = \underline{\mathbf{e}}'_i\mathbf{\Sigma}\underline{\mathbf{e}}_i = \lambda_i \qquad i = 1, 2, ..., p \qquad (2.18)$$

$$Cov(Y_i, Y_k) = \underline{\mathbf{e}}'_i\mathbf{\Sigma}\underline{\mathbf{e}}_k = 0 \qquad i \neq k \qquad (2.19)$$

b) Sendo $Y_1 = \underline{\mathbf{e}}'_1\underline{\mathbf{X}}$, $Y_2 = \underline{\mathbf{e}}'_2\underline{\mathbf{X}}$, ..., $Y_p = \underline{\mathbf{e}}'_p\underline{\mathbf{X}}$ as componentes principais.

Então

$$\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2 = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i) \qquad (2.20)$$

Ou seja, o somatório das variâncias das componentes principais é igual ao somatório das variâncias das variáveis originais.

c) A proporção explicada da variância total pela i-ésima componente principal é dada por:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}, \text{ com i= 1, 2, ..., p} \qquad (2.21)$$

d) Se $Y_1 = \underline{\mathbf{e}}'_1\underline{\mathbf{X}}$, $Y_2 = \underline{\mathbf{e}}'_2\underline{\mathbf{X}}$, ..., $Y_p = \underline{\mathbf{e}}'_p\underline{\mathbf{X}}$ são as componentes principais obtidas da matriz de covariância Σ, então:

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad \text{i, k = 1, 2, ..., p} \qquad (2.22)$$

que é o coeficiente de correlação entre a i-ésima componente principal Y_i e a k-ésima variável X_k.

2.2.2 Componentes principais de variáveis padronizadas

As componentes principais também podem ser obtidas para variáveis padronizadas:

$$\underline{Z}' = \left[Z_1, Z_2, ..., Z_p \right] = \left[\frac{x_1 - \mu_1}{\sqrt{\sigma_1^2}}, \frac{x_2 - \mu_2}{\sqrt{\sigma_2^2}}, ..., \frac{x_p - \mu_p}{\sqrt{\sigma_p^2}} \right] \qquad (2.23)$$

Ou, em notação matricial:

$$\underline{\mathbf{Z}} = \left(V^{1/2} \right)^{-1} \left(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}} \right) \qquad (2.24)$$

onde

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sqrt{\sigma_p^2} \end{bmatrix}, \ \underline{\boldsymbol{\mu}} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \text{ e } \underline{\mathbf{X}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

E pode-se demonstrar que $E(\underline{\mathbf{Z}}) = 0$ e $Cov\left(\underline{\mathbf{Z}}\right) = \left(\mathbf{V}^{1/2}\right)^{-1} \sum \left(\mathbf{V}^{1/2}\right)^{-1} = \boldsymbol{\rho}$. Neste caso, as componentes principais de $\underline{\mathbf{Z}}$ podem se obtidas pelos autovetores da matriz de correlação ρ de $\underline{\mathbf{X}}$.

De acordo com o desenvolvimento de resultados para componentes principais populacionais, obtém-se o desenvolvimento de resultados importantes para componentes principais de variáveis padronizadas:

a) A i-ésima componente principal de variáveis padronizadas $\underline{\mathbf{Z}}' = [Z_1, Z_2, ..., Z_p]$ com $Cov(\underline{\mathbf{Z}}) = \boldsymbol{\rho}$, é dada por:

$$Y_i = \underline{\mathbf{e}}_i' \underline{\mathbf{Z}} = \underline{\mathbf{e}}_i' \left(\mathbf{V}^{1/2}\right)\left(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}}\right), \qquad i = 1, 2, ..., p \qquad (2.25)$$

b) $$\sum_{i=1}^{p} Var\left(Y_i\right) = \sum_{i=1}^{p} Var(Z_i) = p \qquad (2.26)$$

c) O coeficiente de correlação entre a i-ésima componente Y_i e a k-ésima variável padronizada Z_k é dado por:

$$\rho_{Y_i, Z_k} = e_{ki}\sqrt{\lambda_i} \qquad i, k = 1, 2, ..., p \qquad (2.27)$$

d) A proporção da variância total da população (padronizada) explicada pela k-ésima componente principal é dada por

$$\frac{\lambda_k}{p}, \qquad k = 1, 2, ..., p \qquad (2.28)$$

onde λ_k é o k-ésimo autovalor de ρ.

2.2.3 Componentes principais amostrais

Na prática são desconhecidos os valores dos parâmetros $\underline{\boldsymbol{\mu}}$ e $\boldsymbol{\Sigma}$ e, portanto, devem ser estimados. Considerando que $\underline{\mathbf{x}}_1$, $\underline{\mathbf{x}}_2$, ..., $\underline{\mathbf{x}}_p$ são vetores com p x 1 observações independentes de $\underline{\mathbf{X}}$, as estimativas de $\underline{\boldsymbol{\mu}}$ e $\boldsymbol{\Sigma}$ dos vetores de observações independentes de $\underline{\mathbf{X}}$ são respectivamente:

$$\hat{\underline{\boldsymbol{\mu}}} = \bar{\underline{\mathbf{X}}} = \frac{1}{p}\sum_{i=1}^{p} \underline{\mathbf{X}}_i \qquad (2.29)$$

$$\mathbf{S} = \frac{1}{p-1}\sum_{i=1}^{p} (\underline{\mathbf{X}}_i - \bar{\underline{\mathbf{X}}})(\underline{\mathbf{X}}_i - \bar{\underline{\mathbf{X}}})' \qquad (2.30)$$

Assim, a i-ésima componente principal amostral é dada por:

$$\hat{Y}_i = \hat{\underline{\mathbf{e}}}_i' \underline{\mathbf{X}} = \hat{e}_{1i}X_1 + \hat{e}_{2i}X_2 + \ldots + \hat{e}_{pi}X_p, \qquad i = 1, 2, ..., p \qquad (2.31)$$

onde $(\hat{\lambda}_1 \hat{\underline{\mathbf{e}}}_1), (\hat{\lambda}_2 \hat{\underline{\mathbf{e}}}_2), \ldots, (\hat{\lambda}_p \hat{\underline{\mathbf{e}}}_p)$ são os pares de autovalores e autovetores de $\mathbf{S}$ com $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$, com

$$\mathrm{Var}(\hat{Y}_i) = \hat{\lambda}_i, \qquad i = 1, 2, ..., p \qquad (2.32)$$

$$\mathrm{Cov}(\hat{Y}_i, \hat{Y}_k) = 0, \qquad i \neq k \qquad (2.33)$$

Obtêm-se os seguintes resultados, semelhantes aos anteriores mostrados:

a) $$\sum_{i=1}^{p} s_i^2 = \sum_{i=1}^{p} \hat{\lambda}_i \qquad i = 1, 2, ..., p \qquad (2.34)$$

b) A proporção da variância total explicada devido a i-ésima componente principal estimada é dada por:

$$\frac{\hat{\lambda}_i}{\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p} \qquad i = 1, 2, ..., p \qquad (2.35)$$

c) O coeficiente de correlação amostral entre a i-ésima componente principal $\hat{Y}_i$ e a k-ésima variável X_k é dado por:

$$r_{\hat{Y}_i,X_k} = \frac{\hat{e}_{ki}\sqrt{\lambda_i}}{\sqrt{s_k^2}} \qquad i, k = 1, 2, ..., p \qquad (2.36)$$

Para um vetor de observações padronizadas $\hat{\underline{\mathbf{Z}}}' = \left[\hat{Z}_1, \hat{Z}_2, ..., \hat{Z}_p\right]$ a matriz de covariância passa a ser matriz de correlação das variáveis padronizadas. Para obtenção das componentes principais amostrais a partir de variáveis padronizadas, basta seguir o desenvolvimento de componentes principais populacionais para variáveis padronizadas, sendo que os parâmetros serão substituídos pelos seus respectivos estimadores. Deste modo tem-se que (MARQUES, 2006):

a) $\hat{Y}_i = \underline{\mathbf{e}}_i'\underline{\mathbf{Z}} = \hat{e}_{1i}Z_1 + \hat{e}_{2i}Z_2 + \ldots + \hat{e}_{pi}Z_p$ $\qquad i = 1, 2, ..., p \qquad (2.37)$

b) $\mathrm{Var}(\hat{Y}_i) = \hat{\lambda}_i$, $\qquad i = 1, 2, ..., p \qquad (2.38)$

c) $\mathrm{Cov}(\hat{Y}_i, \hat{Y}_k) = 0$ $\qquad i \neq k \qquad (2.39)$

d) Variância total amostral = $\sum_{i=1}^{p} \hat{\lambda}_i = p$ $\qquad (2.40)$

e) $\hat{r}_{\hat{Y}_i,X_k} = \hat{e}_{ki}\sqrt{\hat{\lambda}_i}$ $\qquad i = k = 1, 2, ..., p \qquad (2.41)$

f) A proporção da variância amostral explicada pela i-ésima componente principal é dada por:

$$\frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, ..., p \qquad (2.42)$$

2.2.4 Critérios para determinação do número "k" de componentes principais

Quando a finalidade da aplicação da técnica é a redução da dimensionalidade do espaço amostral, isto é, a sumarização da informação das "p-variáveis" originais em "k" componentes principais, faz-se necessário estabelecer critérios para a seleção do número "k", que é o número de componentes principais a serem retidas no sistema. Geralmente são considerados os seguintes critérios (MARQUES, 2006 e MINGOTI, 2005):

1) *Scree Plot* (CATTELL, 1966): representação gráfica dos autovalores $\hat{\lambda}_i$ da matriz de correlação ou covariância, ordenados em modo decrescente de acordo com a respectiva ordem *i* (Figura 2.3). Por este critério, procura-se no gráfico um "ponto de salto", que estaria representando um decréscimo de importância em relação à variância total. O valor de *k* seria, então, igual ao número de autovalores anteriores ao "ponto de salto". Alguns autores, no entanto, sugerem manter também a primeira componente principal após a formação de cotovelo (CATTEL & JASPERS[5] ,1967, citados por VEGA *et al*, 1998).

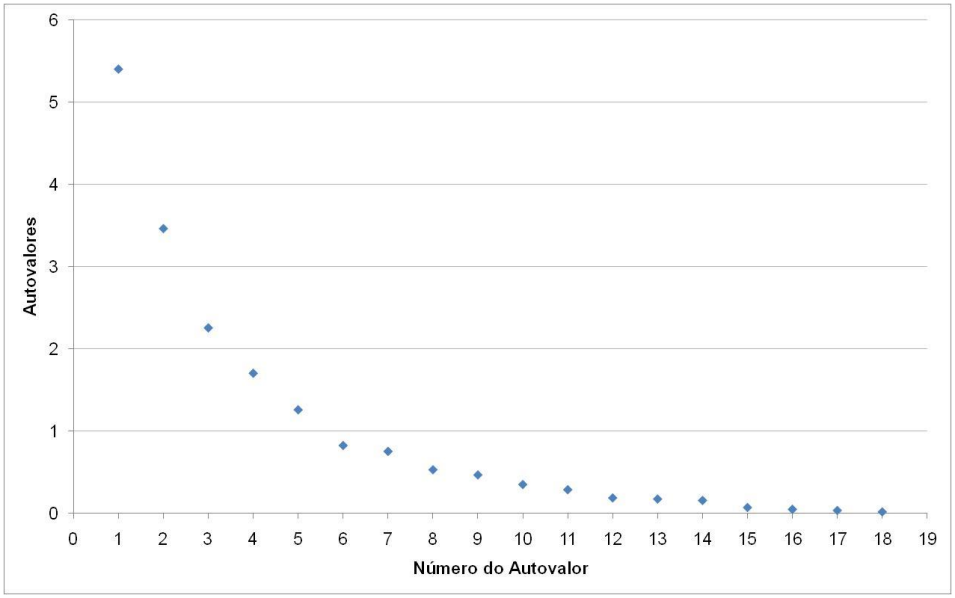


FIGURA 2.3 – Exemplo: *Scree Plot*

2) Análise da representatividade em relação à variância total: de acordo com este critério, deve-se manter no sistema um número de componentes "k" que conjuntamente representem uma porcentagem da variância total. Esta porcentagem da variância total é um valor pré-determinado pelo pesquisador, não havendo um limite definido, podendo ser escolhido de acordo com a natureza do fenômeno investigado. Em algumas situações, é possível obter-se uma porcentagem de explicação de variância total acima de 90% ou 95% com apenas 1 ou 2 componentes, enquanto que em outras, é necessário um número muito maior. Além disso, em alguns casos torna-se necessário trabalhar com porcentagens de explicação abaixo de 90%.

[5] CATTEL, R.B. & JASPERS, J. **A general plasmode (No. 30-10-5-2) for factor analytic exercises and research**. Mult. Behav. Res. Monogr. 67, p. 1 -212, 1967.

3) Critério de Kaiser (1958): o número de componentes retidas deve ser igual ao número de autovalores maiores que 1. A idéia básica do critério é manter no sistema novas dimensões que representem pelo menos a informação de variância de uma variável original.

Por fim, pode-se contar ainda com a própria experiência do pesquisador no assunto.

2.2.5 Escores das Componentes Principais

As componentes principais são variáveis aleatórias que não podem ser medidas diretamente, mas observadas apenas a partir da informação do vetor aleatório *x*. É comum utilizar os escores das componentes para condução de análise estatística de dados ou para simples ordenação (*ranking*) dos elementos amostrais observados com intuito de identificar aqueles que estão com maiores, ou menores, valores globais das componentes (MINGOTI, 2005). Para obter os escores basta aplicar a fórmula matemática da componente aos dados amostrais, ou seja, substituem-se as variáveis nas componentes pelos seus próprios valores. Para cada nova "linha" de dados, um novo escore será calculado para cada uma das componentes principais.

2.3 ANÁLISE FATORIAL

O principal propósito desta análise é reduzir a contribuição de variáveis menos significantes de modo a simplificar ainda mais a estrutura de dados vinda da ACP. Este último propósito pode ser alcançado rotacionando-se os eixos definidos pela ACP, construindo-se novos grupos de variáveis, denominados fatores. Quando ocorre a rotação, diminui-se a contribuição das variáveis com menor significância e aumenta-se a contribuição das que possuem maior significância. A diferença entre componentes principais e fatores é que enquanto as componentes principais são combinações lineares de variáveis de qualidade de água observáveis, os fatores podem incluir variáveis não-observáveis, hipotéticas e "latentes" (WUNDERLIN *et al*., 2001).

A motivação do modelo fatorial decorre da suposição de que variáveis possam ser agrupadas de acordo com suas correlações e que as variáveis dentro de um grupo particular estão altamente correlacionadas entre si, mas muito pouco correlacionadas com variáveis pertencentes a outro grupo. Assim, admite-se que cada grupo de variáveis represente um fator, o qual é responsável pelas correlações observadas (JOHNSON & WICHERN, 1998).

Para estimar os fatores pode-se utilizar o método da verossimilhança ou o método das componentes principais, no entanto, no caso de se utilizar o método da verossimilhança deve ser verificada a normalidade multivariada dos dados observados.

Além disso, certificar-se que os dados são consistentes com a estrutura da análise fatorial é importante. O teste de esfericidade de Bartlett e a medida de adequacidade da amostra de Kaiser-Meyer-Olkin (KMO) são testes empregados para verificar a validade do emprego da Análise Fatorial.

2.3.1 Teste de esfericidade de Bartlett

O teste de esfericidade de Bartlett testa a hipótese de que as variáveis não são correlacionadas na população. A hipótese básica (H_0) diz que a matriz de correlação da população é uma matriz identidade a qual indica que o modelo fatorial é inapropriado. A estatística do teste é dada por:

$$\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] \ln|R| \qquad (2.43)$$

que tem distribuição qui-quadrado com graus de liberdade $\nu = \frac{p(p-1)}{2}$,

onde: n = tamanho da amostra

p = número de variáveis

$|R|$ = determinante da matriz de correlação

No *software* MATLAB este teste é realizado pela função programada **KMO** (Anexo III). O que ocorre nesta função é a comparação do nível de significância (α), denominado “p-valor”, resultante da combinação do valor calculado do χ^2 (qui-quadrado) e dos graus de liberdade (ν) com o valor “0,05”. Assim, quando p-valor < 0,05, a hipótese básica é rejeitada, indicando que os dados são adequados para a análise fatorial.

É importante lembrar que a aplicação do teste de Bartlett requer que as variáveis envolvidas na análise tenham distribuição normal p-variada (MINGOTI, 2005). Na verdade, o não-atendimento a este requisito não implica no total impedimento do uso da análise fatorial, mas sim na possível perda de confiabilidade nos resultados obtidos.

2.3.2 Medida de adequacidade da amostra Kaiser-Meyer-Olkin (KMO)

Alguns autores sugerem que, para que um modelo de análise fatorial possa ser adequadamente ajustado aos dados, é necessário que a matriz de correlação inversa R_{pxp}^{-1} seja próxima da matriz diagonal (RENCHER[6], 2002 citado por MINGOTI, 2005). A medida de adequacidade da amostra KMO é representada por um índice (MAS) que avalia a adequacidade da análise fatorial, sendo calculada por

$$\text{MSA} = \frac{\sum_{j \neq k} \sum_{j \neq k} r_{jk}^2}{\sum_{j \neq k} \sum_{j \neq k} r_{jk}^2 + \sum_{j \neq k} \sum_{j \neq k} q_{jk}^2} \qquad (2.44)$$

onde:

r_{jk}^2 é o quadrado dos elementos da matriz de correlação original (fora da diagonal);

q_{jk}^2 é o quadrado dos elementos fora da diagonal da matriz anti-imagem (onde q_{jk} é o coeficiente de correlação parcial entre as variáveis X_j e X_k).

Conforme Hair Jr, Anderson e Tatham (1987), valores altos - entre 0,5 e 1,0 – indicam que a análise fatorial é apropriada, enquanto que valores baixos, abaixo de 0,5 indicam que a análise fatorial pode ser inadequada. Kaiser e Rice (1978)[7] citados por Sharma (1996) também apresentaram critérios sobre a recomendação da utilização da análise fatorial (Quadro 2.1):

QUADRO 2.1 – Recomendação da aplicação da análise fatorial segundo a medida KMO por Kaiser e Rice (1978)

Medida KMO	**Recomendação**
≥ 0,90	Sensacional
0,80 +	Merecedor
0,70 +	Razoável
0,60 +	Medíocre
0,50 +	Miserável
< 0,50	Inaceitável

[6] RENCHER, A.C. **Methods of multivariate analysis**. New York: John Wiley, 2002.
[7] KAISER, H.F. and RICE, J. **Little Jiffy Mark IV**. Educational and Psychological Measurement, 34 (Spring), p. 111-117, 1974.

2.3.3 Modelo Fatorial Ortogonal

Seja $\underline{\mathbf{X}}$ um vetor aleatório, com média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$. No modelo fatorial $\underline{\mathbf{X}}$ é linearmente independente de algumas variáveis aleatórias não-observáveis F_1, F_2, ..., F_m denominadas fatores comuns e “p” fontes de variações aditivas ε_1, ε_2, ..., ε_p chamadas de erros ou fatores específicos.

O modelo de análise fatorial é:

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p \end{aligned} \qquad (2.45)$$

onde:

μ_i = média da i-ésima variável

ε_i = i-ésimo erro ou fator específico

F_j = j-ésimo fator comum

ℓ_{ij} = Peso ou carregamento na i-ésima variável X_i devido ao j-ésimo fator F_j

$i = 1, 2, \ldots, p$ $\quad$ $j = 1, 2, \ldots, m$ $\quad$ com $\quad m \leq p$

ou, em notação matricial

$$\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}} = \mathbf{L}\underline{\mathbf{F}} + \underline{\boldsymbol{\varepsilon}} \qquad (2.46)$$

$$_{p}\mathbf{L}_{m} = \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1m} \\ \ell_{21} & \ell_{22} & \cdots & \ell_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{p1} & \ell_{p2} & \cdots & \ell_{pm} \end{bmatrix} \qquad _{m}\underline{\mathbf{F}}_{1} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} \qquad _{p}\underline{\boldsymbol{\varepsilon}}_{1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

onde **L** é a matriz de pesos dos fatores.

O que distingue o modelo fatorial do modelo de regressão múltipla é que no modelo de regressão múltipla as variáveis independentes podem ser observadas.

Assumem-se as hipóteses:

$$E(\underline{\mathbf{F}}) = {}_m\underline{\mathbf{0}}_1 \tag{2.47}$$

$$Cov(\underline{\mathbf{F}}) = E(\underline{\mathbf{F}}\underline{\mathbf{F}}') = {}_m\mathbf{I}_m \tag{2.48}$$

$$E(\underline{\boldsymbol{\varepsilon}}) = {}_p\underline{\mathbf{0}}_1 \tag{2.49}$$

$$Cov(\underline{\boldsymbol{\varepsilon}}) = E(\underline{\boldsymbol{\varepsilon}}\underline{\boldsymbol{\varepsilon}}') = {}_p\boldsymbol{\Psi}_p = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{2.50}$$

$$Cov(\underline{\boldsymbol{\varepsilon}}, \underline{\mathbf{F}}) = E(\underline{\boldsymbol{\varepsilon}}\underline{\mathbf{F}}) = {}_p\underline{\mathbf{0}}_m \tag{2.51}$$

As hipóteses descritas e o modelo visto constituem o modelo fatorial ortogonal. Este implica em uma estrutura de covariância para $\underline{\mathbf{X}}$, como segue:

1) $Cov(\underline{\mathbf{X}}) = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$

ou

$$Var(X_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \Psi_i$$

$$Cov(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km} \tag{2.52}$$

2) $Cov(\underline{\mathbf{X}}, \underline{\mathbf{F}}) = \mathbf{L}$

ou

$$Cov(X_i, F_i) = \ell_{ij} \tag{2.53}$$

Conforme Johnson & Wichern (1998, p. 517), as provas de (2.52) e (2.53) são respectivamente:

$$\boldsymbol{\Sigma} = Cov(\underline{\mathbf{X}}) = E(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}})(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}})'$$

$$\boldsymbol{\Sigma} = \mathbf{L}E(\underline{\mathbf{F}}\underline{\mathbf{F}}')\mathbf{L}' + E(\underline{\boldsymbol{\varepsilon}}\underline{\mathbf{F}}')\mathbf{L}' + \mathbf{L}E(\underline{\mathbf{F}}\underline{\boldsymbol{\varepsilon}}') + E(\underline{\boldsymbol{\varepsilon}}\underline{\boldsymbol{\varepsilon}}')$$

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$$

$$Cov(\underline{\mathbf{X}}, \underline{\mathbf{F}}) = E(\underline{\mathbf{X}} - \underline{\boldsymbol{\mu}})\underline{\mathbf{F}}' = \mathbf{L}E(\underline{\mathbf{F}}\underline{\mathbf{F}}') + E(\underline{\boldsymbol{\varepsilon}}\underline{\mathbf{F}}') = \mathbf{L}$$

A porção da variância que a variável contribui para o fator comum “m” é denominada de comunalidade. A porção da variância $\mathrm{Var}(X_i) = \sigma_{ii} = \sigma_i^2$ devido ao fator específico denomina-se especificidade ou variância específica. Tem-se que

$$\underset{\text{Var}(X_i)}{\sigma_{ii}} = \underset{\text{Comunalidade}}{\ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2} + \underset{\text{Variância específica}}{\psi_i} \qquad (2.54)$$

Denotando-se a i-ésima comunalidade por h_i^2 tem-se

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 = \sum_{j=1}^{m} \ell_{ij}^2 \qquad \text{com i = 1, 2, ..., p} \qquad (2.55)$$

então

$$\sigma_i^2 = \sigma_{ii} = h_i^2 + \psi_i, \qquad \text{i = 1, 2, ..., p} \qquad (2.56)$$

A i-ésima comunalidade é a soma dos quadrados dos carregamentos da i-ésima variável com m fatores comuns.

2.3.4 Método das componentes principais para estimar os pesos e as variâncias específicas

Sejam os pares de autovalores-autovetores $(\hat{\lambda}_i, \hat{\underline{e}}_i)$ de **S** (matriz de covariância amostral) com $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ e seja m < p o número de fatores comuns. A matriz dos pesos ou carregamentos estimados dos fatores $\hat{l}_{ij}$ é dada por $\hat{L} = \hat{C}\hat{D}_{\lambda}^{1/2}$. Onde:

$$\hat{C} = \begin{bmatrix} \hat{e}_{11} & \hat{e}_{12} & \cdots & \hat{e}_{1p} \\ \hat{e}_{21} & \hat{e}_{22} & \cdots & \hat{e}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}_{p1} & \hat{e}_{p2} & \cdots & \hat{e}_{pp} \end{bmatrix} \qquad \text{e} \qquad \hat{D}_{\lambda}^{1/2} = \begin{bmatrix} \sqrt{\hat{\lambda}_1} & & & \\ & \sqrt{\hat{\lambda}_2} & & \\ & & \ddots & \\ & & & \sqrt{\hat{\lambda}_p} \end{bmatrix}$$

sendo $\hat{C}$ a matriz dos autovetores dispostos em ordem decrescente e $\hat{D}_{\lambda}^{1/2}$ a matriz diagonal dos autovalores também dispostos em ordem decrescente.

A matriz de pesos ou carregamentos estimada pode ser escrita então do seguinte modo:

$$\tilde{L} = \left[\sqrt{\hat{\lambda}_1}\hat{\underline{e}}_1 \vdots \sqrt{\hat{\lambda}_2}\hat{\underline{e}}_2 \vdots \cdots \vdots \sqrt{\hat{\lambda}_m}\hat{\underline{e}}_m \right] \tag{2.57}$$

No uso deste desenvolvimento para estimar os pesos e as variâncias específicas do conjunto de dados $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_p$, deve-se centralizar as observações. As observações centralizadas têm a forma:

$$\underline{\mathbf{x}} - \underline{\bar{\mathbf{x}}} = \begin{bmatrix} x_{1j} - \bar{x}_1 \\ x_{2j} - \bar{x}_2 \\ \vdots \\ x_{pj} - \bar{x}_p \end{bmatrix}, \quad j = 1, 2, ..., n \tag{2.58}$$

ou padronizando, tem-se

$$\underline{\mathbf{z}}_j = \begin{bmatrix} \dfrac{\left(x_{1j} - \bar{x}_1\right)}{s_1} \\ \dfrac{\left(x_{2j} - \bar{x}_2\right)}{s_2} \\ \vdots \\ \dfrac{\left(x_{pj} - \bar{x}_p\right)}{s_p} \end{bmatrix}, \; j = 1, 2, ..., n \tag{2.59}$$

Neste caso, a matriz de covariância amostral **S** torna-se a matriz de correlação.

As variâncias específicas estimadas são fornecidas pelos elementos diagonais da matriz $\hat{\psi} = \mathbf{S} - \hat{L}\hat{L}'$, assim

$$\hat{\psi} = \begin{bmatrix} \hat{\psi}_1 & & & \\ & \hat{\psi}_2 & & \\ & & \ddots & \\ & & & \hat{\psi}_p \end{bmatrix} \quad \text{com } \hat{\psi}_i = s^2_i - \sum_{j=1}^{m} \hat{l}^2_{ij} \tag{2.60}$$

As comunalidades são estimadas da seguinte forma:

$$\hat{h}^2_i = \hat{l}^2_{i1} + \hat{l}^2_{i2} + \cdots + \hat{l}^2_{im} \quad \text{com i = 1, 2, ..., p} \tag{2.61}$$

2.3.5 Método da máxima verossimilhança para estimar os pesos e as variâncias específicas

Segundo Johnson & Wichern[8] (1998, citados por MARQUES, 2006), para se aplicar a estimação pelo método da máxima verossimilhança do fator de carregamento e da variância específica, assume-se que os fatores comuns $\underline{F}$ e os fatores específicos $\underline{\varepsilon}$ são normalmente distribuídos, assim as observações $\underline{X}_j - \underline{\mu} = \underline{L}\underline{F}_j + \underline{\varepsilon}_i$ também são normalmente distribuídas. Pela verossimilhança a distribuição normal p variada tem a forma

$$N_p\left(\underline{\mu},\Sigma\right) = \left(2\pi\right)^{-\frac{np}{2}} \left|\Sigma\right|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}\left(\underline{x}_j-\bar{\underline{x}}\right)\left(\underline{x}_j-\bar{\underline{x}}\right)'+n\left(\bar{\underline{x}}-\underline{\mu}\right)\left(\bar{\underline{x}}-\underline{\mu}\right)'\right)\right]}$$

$$N_p\left(\underline{\mu},\Sigma\right) = \left(2\pi\right)^{\frac{-(n-1)p}{2}} \left|\Sigma\right|^{-\frac{(n-1)}{2}} e^{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}\left(\underline{x}_j-\bar{\underline{x}}\right)\left(\underline{x}_j-\bar{\underline{x}}\right)'\right)\right]} \qquad (2.62)$$

$$\times\left(2\pi\right)^{-\frac{p}{2}} \left|\Sigma\right|^{-\frac{1}{2}} e^{-\left(\frac{n}{2}\right)\left(\bar{\underline{x}}-\underline{\mu}\right)'\Sigma^{-1}\left(\bar{\underline{x}}-\underline{\mu}\right)}$$

a qual depende de $\mathbf{L}$ e Ψ devido a $\Sigma = LL' + \Psi$. Deseja-se que L seja bem definida pela imposição de uma condição de unicidade dada por

$$L'\Psi^{-1}L = \Delta \quad \text{(Matriz diagonal)} \qquad (2.63)$$

O estimador de máxima verossimilhança de $\mathbf{L}$ e Ψ sujeito à (2.63) deve ser obtido pela maximização numérica de (2.62), a qual se encontra em vários *softwares* estatísticos. Os estimadores de máxima verossimilhança das comunalidades são:

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \cdots + \hat{\ell}_{im}^2 \text{ para } i = 1, 2, ..., p \qquad (2.64)$$

e a proporção da variância total da amostra dada pelo j-ésimo fator é dada por

$$\frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + s_{22} + + s_{pp}} \qquad (2.65)$$

[8] JOHNSON, R. A.; WICHERN. D. W. **Applied Multivariate Statistical Analysis**. 4 Ed. New Jersey: Prentice Hall, 1998.

A prova de (2.62) encontra-se em Johnson e Wichern (1998, p.531).

A proporção da variância total (padronizada) da amostra dada pelo j-ésimo fator é dada por

$$\frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \cdots + \hat{\ell}_{pj}^2}{p} \tag{2.66}$$

2.3.6 Escores fatoriais estimados

Em muitas vezes é interessante conhecer o valor de cada um dos fatores para uma observação individual $\underline{\mathbf{x}} = \left[x_1 \; x_2 \ldots x_p \right]$. Os valores estimados dos fatores comuns denominam-se escores fatoriais.

Segundo MARQUES (2006), Bartlett sugeriu uma metodologia para a estimativa dos valores dos fatores comuns. A metodologia proposta consistia em estimar $\hat{\underline{F}}$ de $\underline{F}$ minimizando a soma dos quadrados dos erros (fatores específicos) dividido pela sua variância recíproca, ou seja,

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \underline{\varepsilon}' \boldsymbol{\Psi}^{-1} \underline{\varepsilon} = \left(\underline{\mathbf{x}} - \underline{\mu} - \mathbf{L}\underline{F} \right)' \boldsymbol{\Psi}^{-1} \left(\underline{\mathbf{x}} - \underline{\mu} - \mathbf{L}\underline{F} \right) \tag{2.67}$$

que tem como solução os escores fatoriais estimados, dados por

$$\hat{\underline{F}} = \left(\hat{\mathbf{L}}' \hat{\boldsymbol{\psi}}^{-1} \hat{\mathbf{L}} \right)^{-1} \hat{\mathbf{L}}' \hat{\boldsymbol{\psi}}^{-1} \left(\underline{\mathbf{x}} - \bar{\underline{\mathbf{x}}} \right) \tag{2.68}$$

Os escores fatoriais estimados para as variáveis padronizadas são dados por

$$\hat{\underline{F}} = \left(\hat{\mathbf{L}}' \hat{\mathbf{L}} \right)^{-1} \hat{\mathbf{L}}' \underline{\mathbf{z}} \tag{2.69}$$

e a matriz de resíduos por

$$\mathbf{R} - \left(\hat{\mathbf{L}}_{\mathbf{z}} \hat{\mathbf{L}}_z' + \hat{\boldsymbol{\psi}}_z \right) = \mathbf{R} - \hat{\mathbf{L}}_{\mathbf{z}} \hat{\mathbf{L}}_z' - \hat{\boldsymbol{\psi}}_z \tag{2.70}$$

Uma das aplicações dos escores fatoriais é na criação de indicadores (escores finais) para classificação. O escore final (E) é dado por

$$E = \hat{F}_1 \times \%\text{variância explicada por } \hat{F}_1 + \hat{F}_2 \times \%\text{variância explicada por } \hat{F}_2 + \ldots + \hat{F}_m \times \%\text{variância explicada por } \hat{F}_m \tag{2.71}$$

2.3.7 Seleção do número de fatores

Para escolha do número de fatores pode-se seguir os mesmos critérios descritos para seleção do número de componentes principais conforme item 2.2.4.

2.3.8 Rotação dos fatores

Sabe-se da álgebra matricial que uma transformação ortogonal corresponde a uma rotação nas coordenadas dos eixos. Essa transformação ortogonal sobre os fatores de carregamento é chamada de rotação dos fatores, que tem como objetivo principal obter pesos altos para cada variável em um único fator e pesos baixos ou moderados nos demais fatores (MARQUES, 2006).

Seja **T** a matriz de transformação e $\hat{\mathbf{L}}$ a matriz estimada dos pesos dos fatores obtidas por qualquer método, então

$$\hat{\mathbf{L}}^{*} = \hat{\mathbf{L}}\mathbf{T}, \text{ onde } \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I} \tag{2.72}$$

é uma matriz de carregamento rotacionada.

Quando um número de fatores é igual a dois (m = 2), pode-se obter graficamente a rotação dos fatores, porém com m > 2 fica impraticável a análise gráfica e torna-se indispensável o uso de programas computacionais para a determinação da rotação dos fatores.

Segundo Marques (2006), Kaiser sugeriu uma medida analítica para efetuar a rotação dos fatores denominada rotação varimax. O procedimento varimax seleciona a transformação ortogonal T que torna

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*4} - \frac{\left(\sum_{i=1}^{p}\tilde{\ell}_{ij}^{*2}\right)}{p}\right] \tag{2.73}$$

o maior possível. Onde $\tilde{\ell}_{ij}^{*}$ são os coeficientes finais rotacionados escalonados pela raiz quadrada das comunalidades, dada por

$$\tilde{\ell}_{ij}^{*} = \frac{\hat{\ell}_{ij}^{*}}{\hat{h}_{i}} \tag{2.74}$$

2.4 ANÁLISE DE AGRUPAMENTOS OU *CLUSTER*

A análise de agrupamentos é uma técnica distinta dos métodos de classificação (análise discriminante, regressão logística). Na classificação tem-se um número de grupos conhecidos e o objetivo é alocar uma nova observação em um destes grupos. Agrupar é uma técnica mais primitiva, no sentido de que nenhuma suposição é feita quanto ao número de grupos, ou estrutura de agrupamento (MARQUES, 2003). Diferentemente da ACP, que normalmente utiliza duas ou três componentes principais para exposição dos seus propósitos, a análise de agrupamentos utiliza toda a variância ou informação contida no conjunto de dados original (VEGA *et al*., 1998).

A análise de agrupamentos, também conhecida como análise de conglomerados, classificação ou *cluster*, tem como objetivo dividir os elementos da amostra, ou população, em grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características (MINGOTI, 2005).

O uso da análise de agrupamentos faz-se presente em diversas situações, não sendo diferente para área ambiental. Em Ecologia, por exemplo, é utilizada na classificação de espécies (McGARIGAL[9] *et al*.,2000, citados por MINGOTI, 2005). Vega *et al*. (1998) utilizaram a análise de agrupamentos para identificar quais amostras de água eram mais homogêneas entre si, com isso obtiveram dois grupos: um que apresentou amostras que sugeriam melhor qualidade da água e o outro com amostras que refletiam pior qualidade de água.

Os critérios a serem utilizados para decidir até que ponto dois elementos do conjunto de dados podem ser considerados semelhantes são as medidas que descrevem a similaridade entre elementos amostrais de acordo com as características neles medidas. Ao considerar que para cada elemento amostral têm-se informações de *p* variáveis armazenadas em um vetor, a comparação de diferentes elementos amostrais poderá ser feita através de medidas matemáticas (métricas), que possibilitam a comparação de vetores, como as medidas de distância. Deste modo, pode-se calcular as distâncias entre os vetores de observações dos elementos amostrais e agrupar aqueles de menor distância.

[9] McGARIGAL, K.; CUSHMAN, S.; STAFFORD,S. **Multivariate statistics for wildlife and ecology research**. New York: Springer Verlag, 2000.

2.4.1 Medidas de similaridade e dissimilaridade

Quando os itens são agrupados, a proximidade é usualmente indicada por um tipo de distância. Já as variáveis são normalmente agrupadas com base nos coeficientes de correlação. Na similaridade quanto maior for o valor observado, mais parecidos são os objetos, como o coeficiente de correlação. Em contrapartida, na dissimilaridade, quanto maior o valor observado menos parecidos eles serão. Existem na literatura várias medidas de dissimilaridade, sendo que cada uma delas produz um determinado tipo de agrupamento. Algumas medidas mais comuns, apropriadas para variáveis quantitativas, são:

Distância Euclidiana:

$$d\left(\underline{x},\underline{y}\right)=\sqrt{\sum_{i=1}^{p}\left(x_i-y_i\right)^2} \tag{2.75}$$

Quadrado da distância Euclidiana:

$$d\left(\underline{x},\underline{y}\right)=\sum_{i=1}^{p}\left(x_i-y_i\right)^2 \tag{2.76}$$

Distância city-block (Manhattan):

$$d\left(\underline{x},\underline{y}\right)=\sum_{i=1}^{p}\left|x_i-y_i\right| \tag{2.77}$$

Distância de Mahalanobis:

$$d\left(\underline{x},\underline{y}\right)=\sqrt{\left(x_i-y_i\right)^{'}S^{-1}}\left(\underline{x}-\underline{y}\right)=\sqrt{\frac{\left(x_1-y_1\right)^2}{s_1^2}+\ldots+\frac{\left(x_p-y_p\right)^2}{s_p^2}} \tag{2.78}$$

2.4.2 Métodos de agrupamentos hierárquicos

As técnicas de agrupamentos hierárquicas são realizadas por série de junções sucessivas tanto como por séries de divisões sucessivas. Os métodos aglomerativos hierárquicos começam com objetos individuais e há inicialmente tantos grupos quanto objetos. Os objetos mais similares são agrupados inicialmente, e esses grupos fundem-se de acordo com suas similaridades. Eventualmente, abrandando o critério de similaridade os sub-grupos vão se unindo a outros sub-grupos até formar um grupo único (MARQUES, 2006).

Johnson & Wichern (1998) definiram um algoritmo de agrupamento aglomerativo hierárquico para N objetos:

1) Inicialmente, há N grupos, cada um contendo um único objeto. Calcula-se a matriz simétrica de distâncias **D = {d_{ik}}**, onde d_{ik} é a distância do objeto i ao objeto k, dada por:

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}, \text{ onde } d_{11} = d_{22} = ... = d_{nn} = 0 \qquad (2.79)$$

2) Encontra-se na matriz simétrica de distâncias D o par de grupos mais próximo, que pode ser representado por dAB, no caso de o grupo A e o grupo B serem os mais próximos. Unem-se estes grupos.

3) Uma nova matriz de distâncias é construída, eliminando-se a coluna e a linha referentes aos grupos A e B formados. Em seguida, adiciona-se uma linha e uma coluna que fornece as distâncias de AB aos outros restantes.

4) Repetem-se os passos dois e três N-1 vezes, observando-se as identidades dos grupos formados e os níveis em que os mesmos se fundem.

O modo de se agrupar os objetos semelhantes é realizado por meio de ligações. Alguns tipos de ligações são: Ligações Simples ou Vizinho mais próximo, Ligações Completas ou Vizinho mais distante, Método das Médias das Distâncias, Método do Centróide e Método de Ward (MARQUES, 2003).

A) Ligações Simples (Vizinho mais próximo)

Neste tipo de ligação, unem-se os dois grupos com menor distância ou maior similaridade. Inicialmente deve-se encontrar a menor distância na matriz simétrica de distância **D = {d_{ik}}** e juntar os objetos correspondentes. Supondo-se que tais objetos sejam o objeto A e o objeto B, representa-se o agrupamento por (AB). Para o passo 3 descrito no item acima, as distâncias entre o grupo (AB) e outro grupo hipotético C é calculado por:

$$d_{(AB)C} = \text{mín}\{d_{AC}, d_{BC}\} \qquad (2.80)$$

As quantidades d_{AC} e d_{BC} são as distâncias entre os vizinhos mais próximos dos grupos A e C e dos grupos B e C.

B) Ligações Completas (Vizinho mais distante)

O procedimento adotado no caso de ligações completas é muito parecido com o caso de ligações simples, diferenciando-se apenas que a distância entre dois grupos é determinada pela distância máxima de dois elementos, uma de cada grupo. Para o passo três do algoritmo proposto, as distâncias entre o grupo (AB) e outro grupo hipotético C é calculado por:

$$d_{(AB)C} = \text{máx}\{d_{AC}, d_{BC}\} \tag{2.81}$$

C) Métodos das Médias das Distâncias

Segundo Mingoti (2005), este método trata a distância entre dois conglomerados (ou grupos) como a média das distâncias entre todos os pares de elementos que podem ser formados com os elemento dos dois conglomerados que estão sendo comparados. Portanto se o grupo 1 tem n_1 elementos e o grupo 2 tem n_2 elementos, a distância entre eles será definida por:

$$d(G_1,G_2)=\sum_{l\in G_1}\sum_{k\in G_2}\left(\frac{1}{n_1 n_2}\right)d(X_l,X_k) \tag{2.82}$$

Assim, a título de exemplificação tem-se a distância entre os grupos $G_1 = \{X_1, X_3, X_7\}$ e $G_2 = \{X_2, X_6\}$ que é igual a:

$$d(G_1,G_2)=\frac{1}{6}\left[d(X_1,X_2)+d(X_1,X_6)+d(X_3,X_2)+d(X_3,X_6)+d(X_7,X_2)+d(X_7,X_6)\right]$$

D) Método do Centróide

Neste método, a distância entre dois grupos é definida como sendo a distância entre os vetores de médias, também chamados centróides, dos grupos que estão sendo comparados. Assim, se $G_1 = \{\mathbf{X_1}, \mathbf{X_3}, \mathbf{X_7}\}$ e $G_2 = \{\mathbf{X_2}, \mathbf{X_6}\}$, por exemplo, os vetores de médias correspondentes são (MINGOTI, 2005):

$$\text{vetor de médias de } G_1=\underline{X_1}=\frac{1}{3}[X_1+X_3+X_7]$$

$$\text{vetor de médias de } G_2 = \underline{X_2} = \frac{1}{2}[X_2 + X_6]$$

e a distância entre G_1 e G_2 é definida por:

$$d(G_1G_2) = \left(\underline{X}_1 - \underline{X}_2\right)' \left(\underline{X}_1 - \underline{X}_2\right) \tag{2.83}$$

que é a distância Euclidiana ao quadrado entre os vetores de médias amostral $\underline{X}_1$ e $\underline{X}_2$. O método do centróide também pode ser utilizado com a distância Euclidiana usual entre os vetores de médias. Em cada passo do algoritmo do agrupamento, os grupos que apresentam o menor valor de distância são agrupados.

O método do centróide é direto e simples. Para fazer o agrupamento, no entanto, em cada passo é necessário voltar-se aos dados originais para o cálculo da matriz de distâncias, o que exige um tempo computacional maior do que nos outros métodos. Ao contrário dos três métodos expostos anteriormente, o método do centróide não pode ser usado em situações nas quais se dispões apenas da matriz de distâncias entre os n elementos amostrais.

E) Método de Ward

Para Mingoti (2005), o procedimento de Ward baseia-se inicialmente na suposição de que cada elemento é considerado um único conglomerado. Em cada passo do algoritmo de agrupamento calcula-se a soma de quadrados dentro de cada conglomerado. Esta soma é o quadrado da distância Euclidiana de cada elemento amostral pertencente ao conglomerado em relação ao correspondente vetor de médias do conglomerado, isto é,

$$SS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \underline{X}_{i.}\right)' \left(X_{ij} - \underline{X}_{i.}\right) \tag{2.84}$$

onde, n_i é o número de elementos no grupo G_i quando se está no passo k do processo e agrupamento, $\mathbf{X}_{ij}$ é o vetor de observação do j-ésimo elemento amostral que pertence ao i-ésimo grupo, $\underline{X}_i$ é o centróide do grupo G_i ,e SS_i representa a soma e quadrados correspondente ao conglomerado G_i . No passo k, a soma de quadrados total dentro dos grupos é definida como:

$$SSR = \sum_{i=1}^{g_k} SS_i \tag{2.85}$$

onde g_k é o número de grupos existentes quando se está no passo k. A distância entre os conglomerados G_l e G_i é, então, definida como:

$$d(G_l,G_i)=\left[\frac{n_l n_i}{n_l+n_i}\right]\left(\underline{X}_{l.}-\underline{X}_{i.}\right)'\left(\underline{X}_{l.}-\underline{X}_{i.}\right) \qquad (2.86)$$

que é a soma de quadrados entres os *clusters* G_l e G_i. Em cada passo do algoritmo de agrupamento, os dois grupos que minimizam a distância (2.86) são combinados.

É possível demonstrar que a medida de distância em (2.86) nada mais é do que a diferença entre o valor SSR depois e antes de se combinar os conglomerados G_l e G_i num único conglomerado. Portanto, em cada passo do agrupamento, o método de Ward combina os dois conglomerados que resultam no menor valor de SSR.

Os resultados obtidos dados o tipo de distância e de ligação são dispostos graficamente em um diagrama em árvore ou dendrograma que possui uma escala para observação dos níveis. A título de exemplo, pode-se observar a Figura 2.4, a qual mostra o agrupamento de 5 objetos individuais iniciais até a formação de um único grupo.

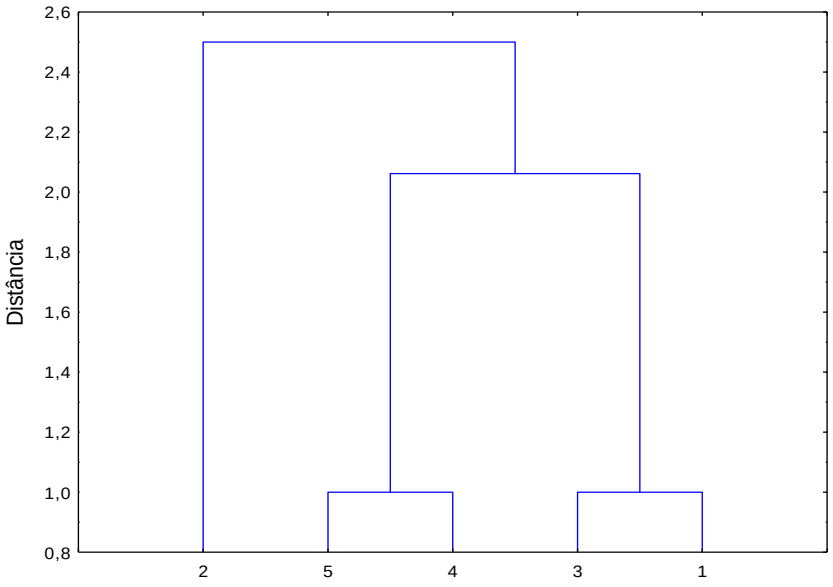


FIGURA 2.4 – Exemplo de dendrograma

2.4.3 Coeficiente de correlação cofenética - Validação do agrupamento

Uma forma de avaliar a validade da informação gerada pela função ligação é compará-la com os dados originais da distância. Se o agrupamento é válido, a ligação dos objetos no agrupamento tem uma forte correlação com as distâncias entre objetos no vetor de distâncias. A função **cofenética** compara esses dois conjuntos de valores e calcula sua correlação. A melhor solução para um agrupamento tem correlação cofenética igual a 1.

O coeficiente de correlação cofenética é calculado utilizando-se a seguinte expressão (CHIGUTI, 2005):

$$r_{cof} = \frac{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(c_{jj'} - \underline{c}\right)\left(f_{jj'} - \underline{f}\right)}{\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(c_{jj'} - \underline{c}\right)}\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(f_{jj'} - \underline{f}\right)}} \qquad (2.87)$$

onde:

$c_{jj'}$ = distância entre as observações j e j' da matriz resultante das ligações que vão ocorrendo

$f_{jj'}$ = distância entre a observação j e j' da matriz de distâncias (item 2.4.1)

$\underline{c} = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(c_{jj'}\right)$, que é a média da matriz *c*

$\underline{f} = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(f_{jj'}\right)$, que é a média da matriz *f*

Sendo *f* a matriz de dissimilaridade (item 2.4.1) e *c* a matriz cofenética resultante da simplificação proporcionada pelo método de agrupamento (ligações).

2.5 APLICAÇÕES DO MÉTODO

Neste item serão apresentados dois estudos de caso que visam elucidar a aplicabilidade da análise multivariada em dados de monitoramento de qualidade de água e exemplificar os diferentes métodos a serem aplicados neste trabalho.

2.5.1 Estudo de Caso 1: Rio Pisuerga, Região Norte da Espanha – Vega *et al*. (1998)

Neste estudo, além da aplicação das análises de componentes principais, fatorial e agrupamentos, utilizou-se a análise de variância ou ANOVA. Foram analisadas 22 variáveis de qualidade de água (Quadro 2.2), coletadas a cada três meses durante dois anos e meio em três estações de monitoramento, resultando em 30 amostras. Desse modo, o requisito para aplicação do método multivariado “n > p”, onde *n* é o número de amostras (ou observações) e *p* é o número de variáveis foi atendido.

QUADRO 2.2 – Parâmetros de qualidade de água do Estudo de Caso 1

Variáveis: Parâmetros de Qualidade da Água	Unidade
Demanda Bioquímica de Oxigênio (DBO)	mg O_2 L^{-1}
Cálcio	mg L^{-1}
Cloreto	mg L^{-1}
Demanda Química de Oxigênio (DQO)	mg O_2L^{-1}
Condutividade	µS cm^{-1}
Sólidos Dissolvidos (SD)	mg L^{-1}
Ferro	mg L^{-1}
Vazão	m^3s^{-1}
Dureza	mg $CaCO_3$ L^{-1}
Bicarbonato	mg L^{-1}
Potássio	mg L^{-1}
Magnésio	mg L^{-1}
Manganês	mg L^{-1}
Sódio	mg L^{-1}
Amônio	mg L^{-1}
Nitrito	mg L^{-1}
Nitrato	mg L^{-1}
Oxigênio Dissolvido (OD)	mg L^{-1}
pH	unidades de pH
Fosfato	mg L^{-1}
Sulfato	mg L^{-1}
Temperatura da Água	°C

Fonte: Adaptado de VEGA *et al*. (1998)

Para evitar erros de classificação em função das diferentes escalas e magnitudes das variáveis de qualidade de água, os autores optaram por normalizar os dados para posteriormente realizarem as análises multivariadas. A Análise Fatorial, realizada a partir da ACP, resultou em 4 fatores, obtidos pelo método *Scree Plot (*CATTELL, 1966*)*. Estes fatores foram responsáveis pela explicação de 67,8% da variância total (Tabela 2.1).

O Fator 1 (F1), que explicou 37,2% da variância total, foi interpretado como Conteúdo Mineral, visto que os parâmetros de qualidade de água cálcio, cloreto, condutividade, sólidos dissolvidos, dureza, magnésio, bicarbonato, sódio e sulfato foram as variáveis com maior peso na definição deste fator (Tabela 2.1).

No Fator 2 (F2) permaneceram as variáveis DBO, DQO e amônia, enquanto que o pH e o OD possuíram uma contribuição negativa para este fator. Neste caso, Vega *et al.* (1998) explicaram que grandes quantidades de matéria orgânica consomem grandes quantidades de oxigênio, assim o sinal negativo do OD expressa esta relação inversa. O peso alto e positivo da amônia deveu-se à decomposição anaeróbia da matéria orgânica. Os autores interpretaram o Fator 2 (F2) como Conteúdo de Matéria Orgânica (poluição antropogênica).

No Fator 3 (F3), a temperatura possuiu um peso alto e positivo, enquanto que o OD possuiu um peso negativo, o que é explicado em razão de a solubilidade dos gases na água diminuir com o aumento da temperatura. Os autores esperavam, ainda, um peso alto e negativo para a vazão no F3, visto que altas temperaturas correspondem à estiagem e ao período de verão, quando a vazão é mais baixa. No entanto, o peso apesar de negativo é baixo, o que ocorreu em virtude de um longo período de estiagem que persistiu inclusive no inverno.

O Fator 4 (F4), que explicou somente 5,9% da variância total, teve a contribuição do ferro e do manganês, que são relacionados “hidroquimicamente”.

A Tabela 2.1 mostra as correlações (pesos) entre os fatores e as variáveis de qualidade de água. Contudo, Vega *et al.*(2008) não estipularam – ou não se citou em texto - um critério quanto à magnitude dos pesos. Ou seja, um valor mínimo que estipula o “corte” de variáveis não tão relevantes para definição do respectivo fator.

TABELA 2.1 – Peso das variáveis em cada um dos fatores

Variável	Fator 1	Fator 2	Fator 3	Fator 4
DBO	0.116	0.934	0.163	0.111
Cálcio	0.920	-0.179	-0.093	-0.119
Cloreto	0.893	0.326	0.048	-0.034
DQO	0.180	0.912	0.159	0.011
Condutividade	0.973	0.148	0.049	-0.038
SD	0.950	0.183	-0.001	0.001
Ferro	-0.131	0.072	0.012	0.970
Vazão	-0.496	-0.005	-0.323	-0.094
Dureza	0.952	0.089	0.106	-0.033
Bicarbonato	0.697	0.184	0.024	-0.139
Potássio	0.584	0.614	0.089	-0.043
Magnésio	0.766	0.359	0.289	0.071
Manganês	0.248	0.290	0.387	0.472
Sódio	0.918	0.180	-0.070	0.003
Amônio	0.225	0.761	-0.190	0.065
Nitrito	0.105	0.170	0.182	-0.061
Nitrato	0.014	-0.003	-0.260	0.104
OD	-0.132	-0.418	-0.540	-0.016
pH	0.169	-0.434	-0.018	-0.201
Fosfato	0.276	0.350	0.244	0.045
Sulfato	0.981	0.008	0.059	0.022
Temperatura	-0.003	0.114	0.919	0.031
% Var. Explicada	37.2	16.7	8	5.9
%Var. Acumulada	37.2	53.9	61.9	67.8

Fonte: Adaptado de VEGA *et al*. (1998)

A Figura 2.5 mostra os escores das amostras do rio no plano definido pelos fatores 1 e 2, onde a ordenada é representada pelo F2 (conteúdo de matéria orgânica) e a abscissa pelo F1 (conteúdo mineral). As amostras são representadas graficamente por um código formado

pela combinação da estação de monitoramento, do mês e do ano em que a amostra foi coletada. A estação de monitoramento *Cabezón* (●) corresponde à seção onde o rio ainda não recebeu efluentes industriais e municipais, mas há poluição difusa; *Puente Mayor* (■) reflete a situação onde há despejos industriais, mas não municipais e *Simancas* (▲) onde o rio já recebeu toda a poluição. As amostragens ocorreram nos meses de Janeiro(E), Abril(A), Julho(J) e Outubro(O) de 1990 a 1992. Assim, a título de exemplificação, o código ■J90 representa uma amostragem ocorrida em *Puente Mayor* no mês de Julho no ano de 1990.

Escores altos e positivos no F1 e F2 indicam grande conteúdo mineral e grande poluição orgânica, respectivamente. Enquanto que amostras com escores altos e negativos nos fatores 1 e 2 correspondem a vazões altas – que contribuem para a diluição dos minerais dissolvidos - e alta concentração de OD, indicando melhor qualidade da água. Assim, a partir da Figura 2.5, pode-se concluir que a amostra ▲J90 apresenta a pior qualidade, com escores altos em F1 e F2.

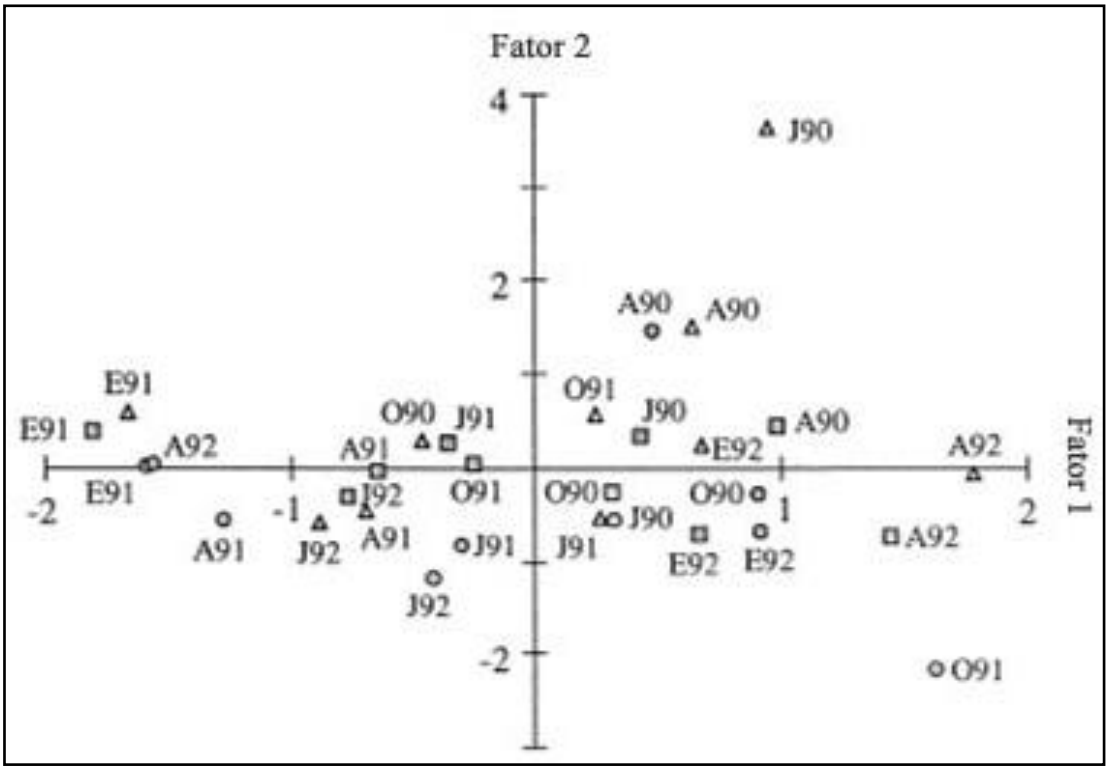


FIGURA 2.5 – Escores das amostras do Rio Pisuerga no plano definido pelos fatores 1 e 2
Fonte: Adaptado de VEGA *et al.* (1998)

A Análise de Agrupamentos ou *Cluster* permitiu o agrupamento das amostras da água do rio baseada nas semelhanças de suas composições químicas. As amostras foram coletadas nas estações *Cabezón* (C), *Puente Mayor* (P) e *Simancas* (S), nos meses de Janeiro (E), Abril

(A), Julho (J) e Outubro (O) de 1990 a 1992. Quanto à legenda das amostras, PE91, por exemplo, é a amostra coletada em *Puente Mayor* em Janeiro/91. Neste estudo, utilizou-se o agrupamento hierárquico, quadrado da distância euclidiana e o método de ligação Ward, visto que este possui um pequeno efeito de distorção em função do espaço e utiliza mais informações que os outros métodos (Willet[10] citado por VEGA, 1998).

Observa-se na Figura 2.6 a formação de agrupamentos, cada um deles formado por dois subgrupos, com a qualidade da água piorando do topo para base.

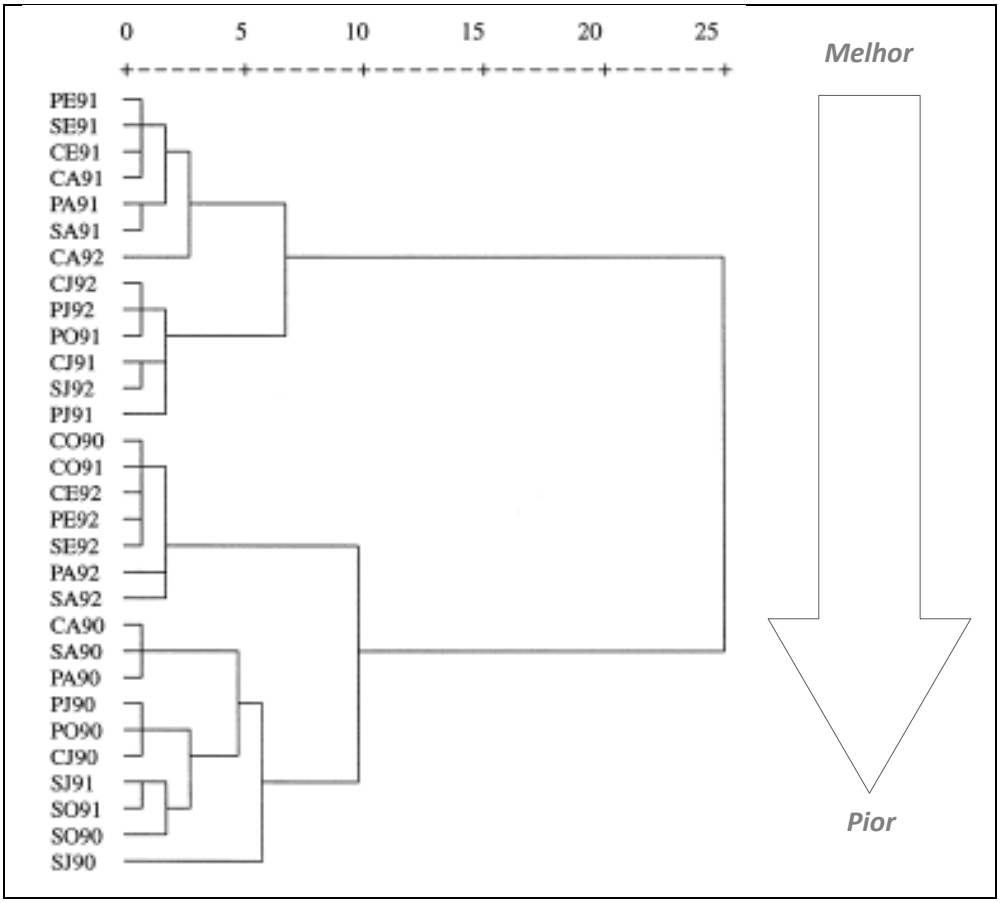


FIGURA 2.6 – Dendrograma referente às amostras coletadas em Cabezón, Puente Mayor e Simancas
Fonte: Adaptado de VEGA *et al.* (1998)

Estes agrupamentos foram interpretados do seguinte modo pelos autores:

[10] WILLET, P. **Similarity and Clustering in Chemical Information Systems**. Research Studies Press, Wiley, New York, 1987.

QUADRO 2.3 – Interpretação dos resultados do Estudo de Caso 1

Grupos	Interpretação
PE91, SE91, CE91, CA91, PA91, SA91, CA92	Baixo conteúdo mineral e de matéria orgânica, evidenciando uma qualidade da água melhor.
CJ92, PJ92, PO91, CJ91, SJ92, PJ91	Na Análise Fatorial, estas amostras apresentaram valores intermediários e negativos no eixo do F1, indicando a presença de conteúdo mineral.
CO90, CO92, CE92, PE92, SE92, PA92, SA92	Estas amostras apresentaram valores altos e positivos no eixo do F1 e negativos no F2, indicando grande conteúdo mineral e pouca matéria orgânica.
CA90, SA90, PA90, PJ90, PO90, CJ90, SJ90, SO91, SO90, SJ90	Estas amostras correspondem à estiagem e à estação de monitoramento mais contaminada (*Simancas*) e mostram a pior qualidade tanto quanto ao conteúdo mineral como orgânico.

Fonte: Adaptado de VEGA *et al.* (1998)

Vega *et al*. (1998) concluíram que os métodos multivariados permitiram a identificação e avaliação das fontes espaciais e temporais de variação que afetam a qualidade e a hidroquímica do corpo hídrico. Demonstrou-se que a poluição orgânica origina-se dos efluentes municipais despejados no rio entre as estações de *Puente Mayor* e *Simancas* e que os efeitos temporais estão associados a variações na vazão que ocasiona a diluição de poluentes e, portanto, variações na qualidade da água. Além disso, a Análise de Agrupamentos gerou uma classificação significativa das amostras do rio, identificando quão poluídas eram.

2.5.2 Estudo de Caso 2: Rio St. Johns, Flórida, Estados Unidos – Ouyang (2005)

Neste estudo, as variáveis foram as estações de monitoramento e não as variáveis de qualidade de água. Foram avaliadas 22 estações de monitoramento, contando para isto com dados de 42 variáveis de qualidade de água monitoradas por 3 anos (1999-2001). Foram utilizados os valores de mediana de cada variável em vez de suas médias em virtude de os dados serem distribuídos de modo desorganizado. O período de 3 anos foi selecionado porque não havia um conjunto de dados completo que incluísse todas as variáveis de qualidade de água além de 3 anos, visto que a ACP requer que não haja valores ausentes na matriz de dados. A Tabela 2.2 mostra 4 das 22 estações de monitoramento e os valores correspondentes a 4 das 42 variáveis de qualidade de água, considerando a mediana.

TABELA 2.2 – Dados de qualidade de água referentes a 4 estações de monitoramento

Parâmetros de Qualidade de Água	Estações de Monitoramento			
	SJR01	SJR04	SJR09	SJR14
Temperatura da Água (°C)	30,18	30,60	16,99	30,83
DBO (mg/L)	0,90	0,95	1,10	1,10
OD (mg/L)	4,98	4,36	8,32	4,01
Turbidez (NTU)	3,07	4,72	6,23	6,23

Fonte: Adaptado de OUYANG (2005)

Ouyang (2005) utilizou o *software Statistica Analysis System* (versão 8) para aplicação da ACP/AF, objetivando identificar quais estações de monitoramento eram realmente importantes para avaliação anual das variações da qualidade do rio. O autor optou por padronizar os dados e, então, utilizar a matriz de covariância para obtenção dos autovalores e autovetores. Na AF, considerou as estações de monitoramento com coeficiente de correlação (peso) maior que 0,75 nos fatores. Assim, estações que apresentaram correlações inferiores a esse valor foram consideradas estações *não-principais*.

A ACP resultou em duas componentes principais que juntas representaram 99,1% da variância total, contudo não foi possível obter qualquer informação sobre quais estações de monitoramento explicavam a maioria da variância. Assim, aplicou-se a análise fatorial para identificar afinal quais estações eram mais importantes quanto às variações da qualidade da água. Neste estudo, o critério (autovalores) utilizado para reter os fatores mais importantes foi o *default* do *software Statistica* ($>10^{-6}$) , o que resultou em 14 fatores. Os fatores 1 e 2 explicaram 99,1% da variância total assim como na ACP, sendo possível visualizar que 3 das 22 apresentavam-se menos importantes, sendo consideradas estações não-principais. O autor comparou então os dados de qualidade com e sem a presença das três estações não-principais, ou seja, considerando as 22 estações e as 19 estações principais, respectivamente. Para isto plotou os dados de Carbono Orgânico Dissolvido *vs*. Cor; Clorofila a *vs*. Fósforo Total; DBO *vs*. Carbono Orgânico Total e Clorofila a *vs*. Nitrogênio Total Dissolvido, ajustando linhas de tendência. Obteve-se para todos os casos que o coeficiente de correlação, R^2, das curvas ajustadas foi maior considerando-se apenas as 19 estações principais. A Figura 2.7 mostra a relação Carbono Orgânico Dissolvido *vs*. Cor, sendo possível observar que o R^2 das 22 estações foi menor do que o referente às 19 estações: 0,5103 < 0,5553.

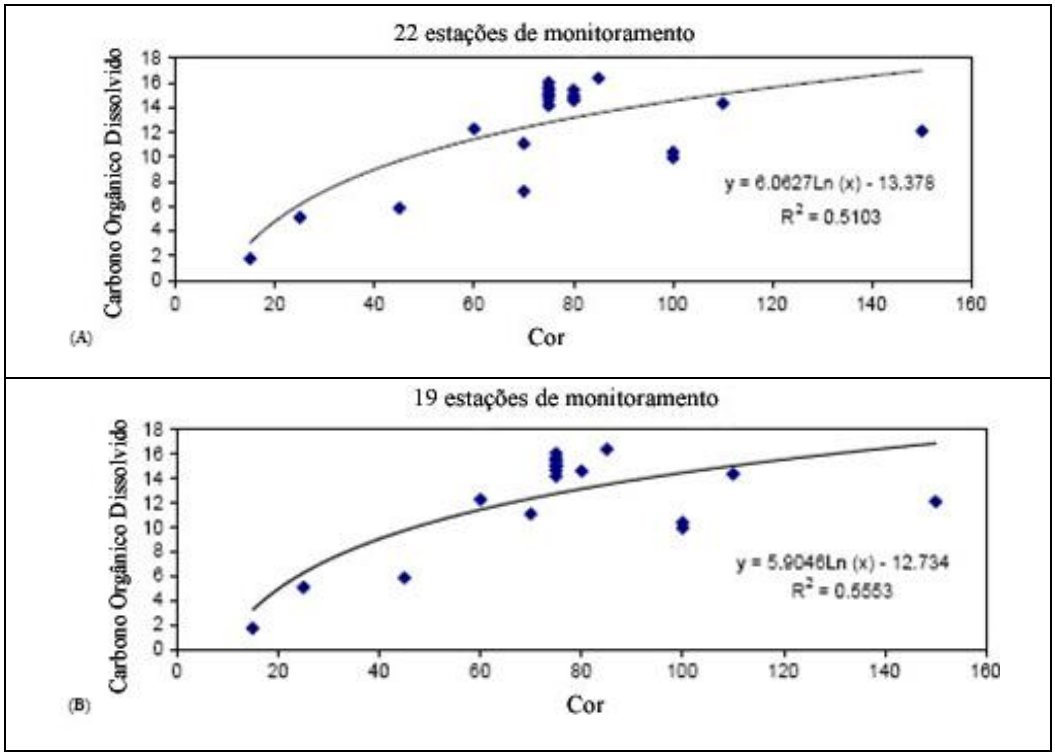


FIGURA 2.7 – Comparação entre as 22 estações de monitoramento (A) e as 19 principais (B), considerando "Cor *vs.* Carbono Orgânico Dissolvido"
Fonte: Adaptado de OUYANG, 2005

Ouyang (2005) concluiu, então, que pode haver uma melhoria na eficiência das estações de monitoramento bem como redução de custos, diminuindo-se o número de estações, sem sacrificar dados importantes de qualidade de água. Contudo, alertou que a decisão real sobre a eliminação de estações de monitoramento deve ser tomada considerando análises de dados pertencentes a períodos mais longos, isto é, mais de três anos.

2.6 SÍNTESE DO CAPÍTULO

Neste capítulo foram abordadas as técnicas multivariadas das Componentes Principais, Fatorial e Agrupamentos, apresentando-se suas bases teóricas. Buscou-se inserir a análise multivariada no contexto da área de gestão de qualidade de água, utilizando-se de experimentos realizados por outros autores como referência para o desenvolvimento desta pesquisa na Bacia do Alto Iguaçu na RMC. Ressalte-se, ainda, que esta se trata da primeira aplicação de técnicas multivariadas com enfoque na gestão de recursos hídricos na Bacia do Alto Iguaçu. Na sequência, são apresentadas sistematizações desenvolvidas neste trabalho que mostram de modo simplificado os métodos estatísticos adotados, o que normalmente não é apresentado na literatura de modo evidente.

i) Sistematização simplificada da Análise de Componentes Principais (ACP)

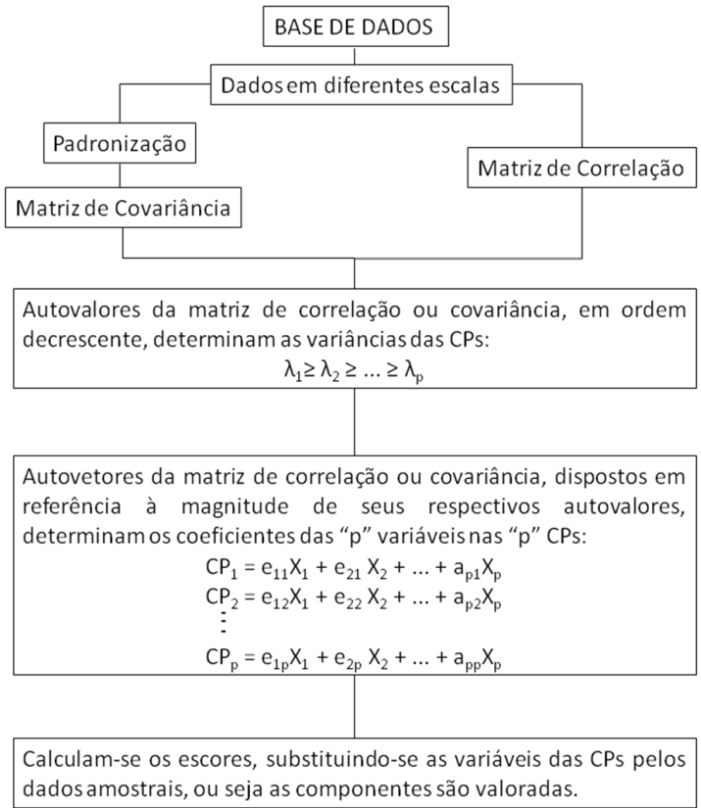
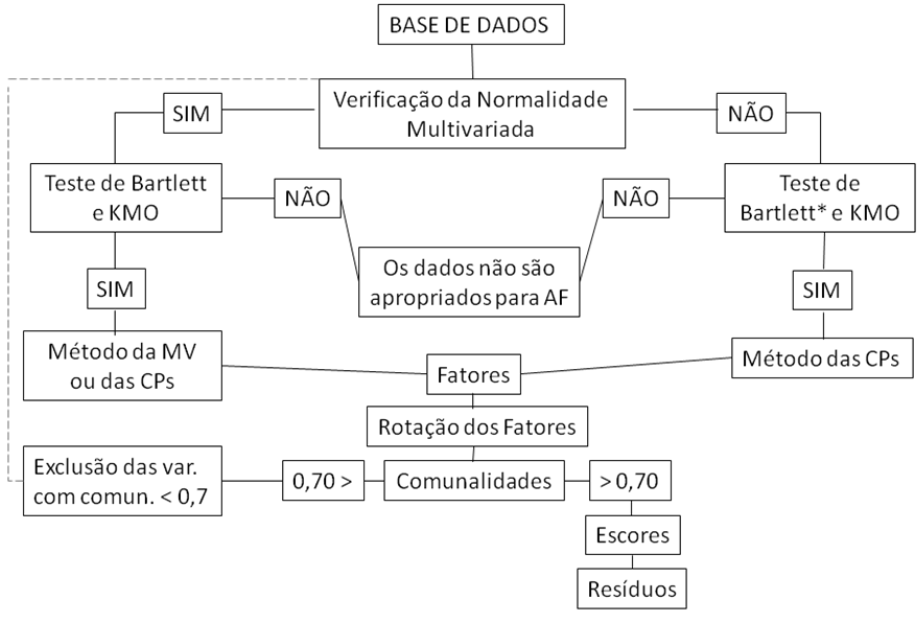


FIGURA 2.8 – Sistematização da Análise de Componentes Principais

ii) Sistematização simplificada da Análise Fatorial (AF)



* Perda de confiabilidade nos resultados, visto que a aplicação do teste de Bartlett requer que as variáveis envolvidas na análise tenham distribuição normal variada

FIGURA 2.9 - Sistematização da Análise Fatorial

iii) Sistematização simplificada da Análise de Agrupamentos

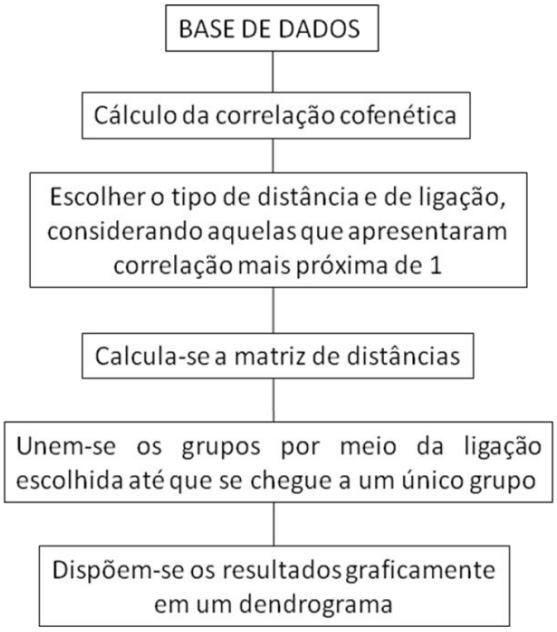


FIGURA 2.10 - Sistematização da Análise de Agrupamentos

CAPÍTULO III

3. AVALIAÇÃO DA QUALIDADE DA ÁGUA EM BACIAS: ESTRATÉGIA PARA AVALIAÇÃO ESTATÍSTICA

Neste capítulo, apresentam-se a bacia do Alto Iguaçu e os pontos de monitoramento localizados em sua extensão, bem como os parâmetros utilizados nesta pesquisa para a avaliação qualitativa e quantitativa do corpo hídrico. A aplicação propriamente dita dos métodos expostos no capítulo anterior e as estratégias utilizadas para a avaliação dos dados monitorados também são apresentados no atual capítulo.

3.1 CARACTERIZAÇÃO DA ÁREA DE ESTUDO

A área de estudo selecionada para avaliação foi a Bacia do Alto Iguaçu (Figura 3.1), situada na região leste do Estado do Paraná, na Região Metropolitana de Curitiba. A bacia é constituída por 26 sub-bacias principais e possui área de drenagem de aproximadamente 2.800 km². O rio Iguaçu possui extensão de cerca de 90 km sendo formado pela junção dos rios Iraí, Iraizinho, Piraquara, Palmital e Atuba (PORTO *et al*., 2007).

Na região situada mais a leste da bacia do Alto Iguaçu estão as nascentes dos rios Iraí, Iraizinho, Piraquara e Pequeno. Trata-se de uma região que faz divisa com a Serra do Mar, onde existem áreas de proteção ambiental, portanto, com menor densidade populacional e mais preservada. Os rios desta região são formadores de represas para o abastecimento público como a represa do Iraí e Piraquara. Os rios Itaqui e Pequeno contribuem para a vazão do canal de água limpa, situado na margem esquerda do rio Iguaçu, com início na soleira da ponte PR-415 (ponto de monitoramento P1) sobre o rio Iraí até o rio Pequeno.

Os principais afluentes do rio Iguaçu pela margem esquerda em seu trecho de montante são os rios: Itaqui, Pequeno, Miringuava, Cotia, Despique, Maurício e Faxinal. Estes apresentam características de qualidade de água mais preservadas ou com menores cargas de poluição que os da margem direita.

Nas regiões situadas na margem direita do Alto Iguaçu, em seu trecho de montante e médio, estão os rios Bacacheri, Belém, Padilha e Barigüi. Estes rios cortam a cidade de Curitiba e recebem toda a carga proveniente da poluição difusa, efluentes doméstico, lançamentos pontuais e efluentes resultantes das indústrias instaladas na CIC, situadas no terço inferior da

bacia do Barigüi. Esta é a região onde estão localizados os rios que mais contribuem para a poluição do rio Iguaçu e que requer maiores controles e monitoramento dos recursos hídricos.

E por fim, o rio Iguaçu a partir de seu terço final, após receber a afluência dos rios Verde e Itaqui pela margem direita, começa a apresentar melhores condições de qualidade da água. Isto se dá em função da autodepuração e pelo fato de receber afluências de melhor qualidade, diluindo assim a sua carga inicial.

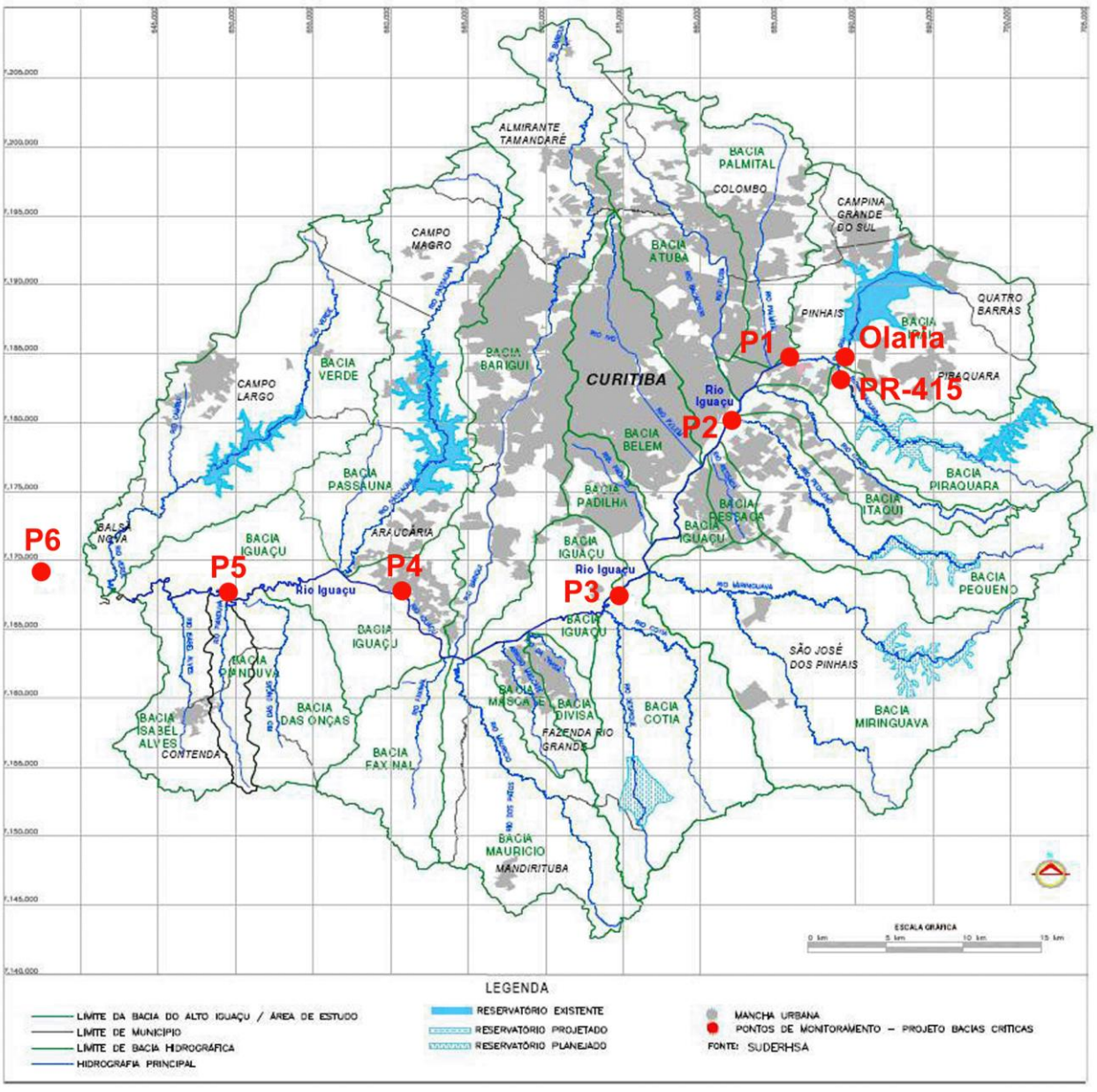


FIGURA 3.1 – Mapa da Bacia do Alto Iguaçu com suas principais sub-bacias
FONTE: KNAPIK *et al*., 2008

A Figura 3.2 apresenta a bacia em forma de diagrama topológico, visando um melhor entendimento de onde se situam as entradas de afluentes e de efluentes de estações de tratamento de esgoto e indústrias.

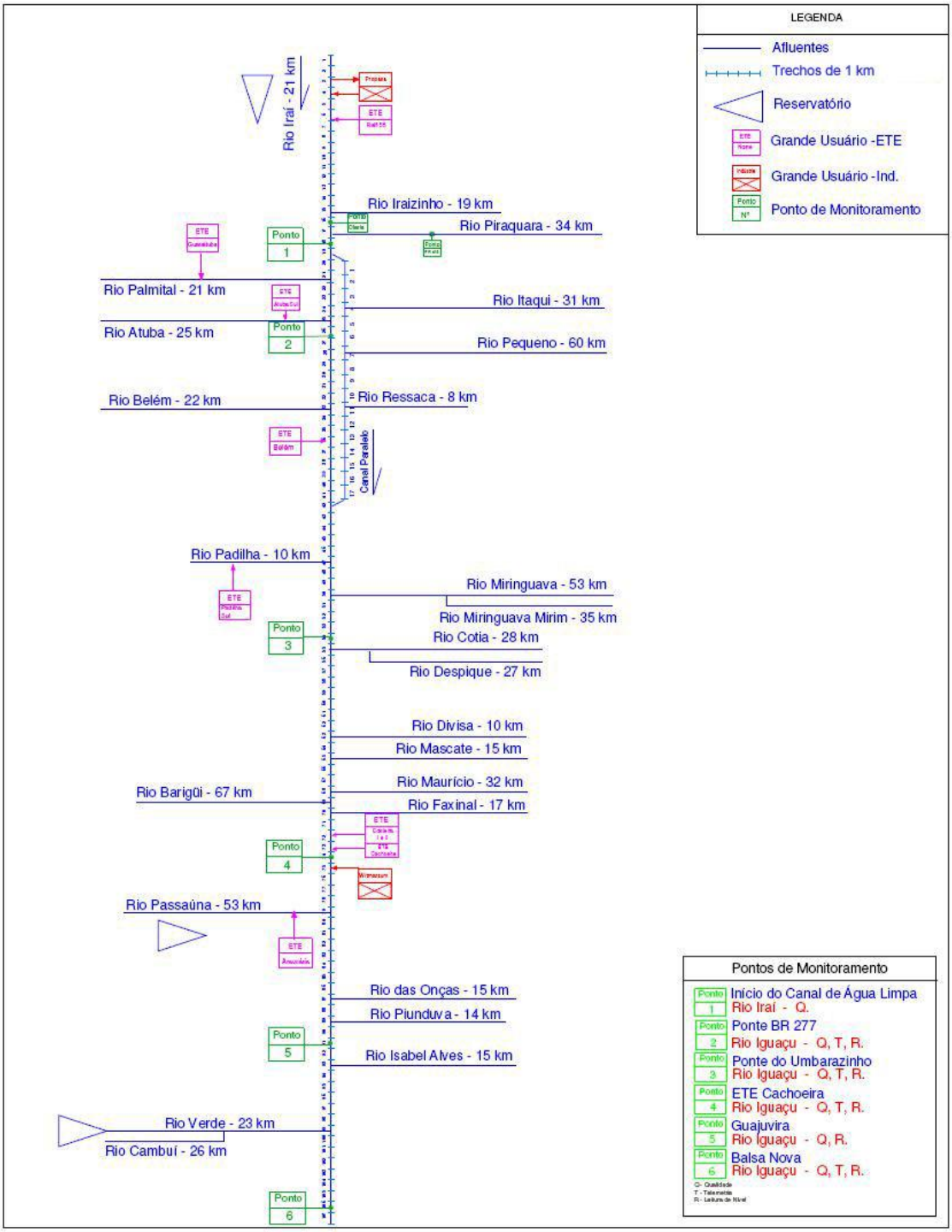


FIGURA 3.2- Diagrama topológico da Bacia do Alto Iguaçu
Fonte: Adaptado de KNAPIK *et al*.(2008)

3.1.1 Aspectos Demográficos

A população existente na bacia é de aproximadamente 3 milhões de habitantes - com cerca de 92% da população total caracterizada como urbana - distribuídos em 14 municípios. No Quadro 3.1, são apresentados os dados populacionais para os municípios localizados na área de estudo, segundo atualização da Coordenação da Região Metropolitana de Curitiba - COMEC, para o ano de 2005 (PORTO *et al*., 2007).

Quadro 3.1 – População estimada para o ano de 2005

Município	**População para 2005 (habitantes)**
Almirante Tamandaré	108.168
Araucária	107.926
Campina Grande do Sul	36.291
Campo Largo	90.044
Campo Magro	25.885
Colombo	215.955
Contenda	7.256
Curitiba	1.735.401
Fazenda Rio Grande	83.800
Mandirituba	7.491
Pinhais	116.824
Piraquara	80.390
Quatro Barras	18.995
São José dos Pinhais	230.144
Total	*2.864.570*

A bacia contemplada é uma região altamente urbanizada, concentrando 25% da população total e 30% da população urbana do estado, e vem passando por um processo de ocupação irregular de várzeas e áreas de mananciais, em especial na margem direita do Rio Iguaçu. Como conseqüência deste processo, têm sido constatados problemas relacionados aos sistemas de abastecimento de água, do tratamento de esgotos sanitários e dos sistemas de drenagem urbana, os quais não acompanham o crescimento das cidades, afetando negativamente o meio ambiente e a qualidade de vida das pessoas.

Nas últimas décadas, de acordo com o estudo desenvolvido pela SUDERHSA (2000), observou-se uma intensificação na tendência de expansão da malha urbana de Curitiba em

direção aos municípios limítrofes, como Fazenda Rio Grande, São José dos Pinhais, Colombo, Almirante Tamandaré, Colombo e Araucária, com a ampliação e o adensamento da urbanização existente, tendendo à integração das diversas sedes municipais.

3.1.2 Aspectos Físicos

A área de estudo, em especial o trecho atravessado pelos rios Iraí e Iguaçu na região metropolitana de Curitiba, é uma região predominantemente plana, apresentando uma grande extensão de várzeas naturais em ambas as margens, configurando planícies de inundação bem definidas. Estas várzeas são locais com solos permanentemente úmidos, com o nível do lençol freático próximo da superfície do terreno, em grande parte coberto por vegetação rasteira típica. Há também uma intensa atividade de extração de areia nas cavas existentes nas áreas mais planas dessas várzeas inundáveis.

3.1.3 Aspectos Climáticos

O clima para a região do Alto Iguaçu, segundo a classificação de Köppen é Cfb. O tipo climático Cfb indica clima mesotérmico ou subtropical, com precipitação média anual de 1.400 mm, com temperatura mínima média de 12,5°C e temperatura máxima média de 22,5°C, estando sujeito a geadas severas.

3.1.4 Atividade Industrial

De acordo com o Cadastro de Usuários, realizado pelo Plano de Despoluição Hídrica da Bacia do Alto Iguaçu (SUDERHSA, 2000), das indústrias cadastradas, 49 contribuem com cerca de 95% da carga de DBO de origem industrial lançada nos rios, solo ou rede de esgoto da bacia do Alto Iguaçu, correspondendo a 57 t DBO/mês, ou o equivalente a uma população de pouco mais de 35.000 habitantes. Em termos de proporcionalidade, a contribuição industrial de dois meses equivale ao esgoto doméstico bruto de um dia da população da bacia do Alto Iguaçu. A maior parte das indústrias cadastradas está localizada nas bacias dos rios Barigüi, Belém, Padilha, Passaúna e Atuba.

3.2 PONTOS DE MONITORAMENTO

A atividade de monitoramento em uma bacia hidrográfica é fundamental para a formação de uma base de dados que permita um melhor conhecimento do corpo hídrico e uma adequada gestão dos recursos hídricos.

Neste trabalho, utilizou-se a mesma base de dados gerada pelo monitoramento da bacia do Alto Iguaçu no âmbito do Projeto Bacias Críticas (PORTO *et al.*, 2007), e, adicionalmente optou-se por complementar a base de dados com dados gerados no ano de 2008 visando à obtenção de um conjunto maior de dados.

Na bacia em estudo, os pontos de monitoramento foram escolhidos baseados nos seguintes critérios: existência de estação fluviométrica (telemetria e/ou régua de nível), fácil acesso para coleta de amostras (proximidade de estradas, pontes) e distância entre pontos consecutivos (em média 20 km para cobrir toda a área de estudo) (PORTO *et al.*, 2007).

Inicialmente, foram selecionados 5 pontos de monitoramento e 2 pontos de visitação para leitura de nível dos afluentes Iraí e Piraquara. Isto porque, no ponto de monitoramento P1 não há estação de telemetria e/ou régua de nível, sendo necessário estimar a vazão em função das vazões das estações de monitoramento Olaria e PR-415. No decorrer das campanhas foi selecionado mais um ponto de monitoramento, localizado a jusante da estação de monitoramento Ponte do Guajuvira (P5), o ponto P6. No início de 2008, optou-se por começar a analisar as amostras também em laboratório para os dados da Olaria, além dos parâmetros já monitorados *in situ*. Os pontos de monitoramento selecionados são apresentados no Quadro 3.2:

QUADRO 3.2 – Pontos de monitoramento na Bacia do Alto Iguaçu

Identificação	**Tipo**	**Localização**	**Rio**
Olaria	Qualidade e Quantidade	Olaria do Estado	Iraí
PR-415	Quantidade	Ponte PR-415	Piraquara
P1	Qualidade	Início do Canal de Água Limpa	Iraí
P2	Qualidade e Quantidade	Ponte BR-277	Iguaçu
P3	Qualidade e Quantidade	Umbarazinho	Iguaçu
P4	Qualidade e Quantidade	ETE Araucária	Iguaçu
P5	Qualidade e Quantidade	Ponte do Guajuvira	Iguaçu
P6	Qualidade e Quantidade	Balsa Nova	Iguaçu

A localização dos pontos de monitoramento pode ser observada na Figura 3.1, os pontos em vermelho indicam as estações Olaria, PR-415, P1, P2, P3, P4, P5 e P6 da direita para a esquerda do mapa.

3.3 ATIVIDADES DE CAMPO

As atividades de campo foram realizadas em dois períodos: um iniciado em junho de 2005 com término em julho de 2006 com freqüência quinzenal e o outro em 2008, começado em março e finalizado em agosto, com freqüência mensal. Para o primeiro período foram realizadas 19 campanhas e para o segundo 5. O Apêndice I apresenta algumas fotos dos pontos monitorados da bacia. O Quadro 3.3 mostra o número total de campanhas, evidenciando em quantas delas os parâmetros eram analisados *in situ* e em laboratório.

QUADRO 3.3 – Número de campanhas realizadas nos pontos de monitoramento

Ponto de Monitoramento	**Nº de campanhas com parâmetros analisados *in situ***	**Nº de campanhas com parâmetros analisados em laboratório**	**Total de campanhas realizadas**
Olaria	21	4	21
P1	24	24	24
P2	24	24	24
P3	24	24	24
P4	24	24	24
P5	23	23	23
P6	21	21	21

3.4 PARÂMETROS DE QUALIDADE DE ÁGUA MONITORADOS

Com o objetivo de se conhecer melhor a dinâmica da bacia do Alto Iguaçu foram realizadas 24 campanhas de monitoramento ao longo de 107 km - considerando além dos 86 km do rio Iguaçu os 21 km do rio Iraí - durante um período de 17 meses não consecutivos, iniciado no ano de 2005 e finalizado no ano de 2008.

Foram monitorados *in situ* 7 parâmetros sendo eles: OD, turbidez, condutividade, temperatura da água, pH, profundidade Secchi e leitura de nível para encontrar posteriormente o valor de vazão através da curva-chave. Em laboratório, foram analisados os seguintes

parâmetros: DBO_5, DQO, COT, série de nitrogênio, fósforo total e sólidos. A descrição dos parâmetros monitorados pode ser acompanhada na sequência.

A. Demanda Bioquímica de Oxigênio - DBO_5

É a quantidade de oxigênio necessária para oxidar a matéria orgânica por decomposição microbiana aeróbia para a forma inorgânica estável. A DBO_5 é normalmente considerada como a quantidade de oxigênio consumida durante um período de 5 dias numa temperatura de incubação de 20°C. Despejos de origem predominantemente orgânica proporcionam os maiores aumentos em termos de DBO num corpo d'água. A presença de um alto teor de matéria orgânica pode induzir à completa extinção do oxigênio na água, provocando o desaparecimento de peixes e outras formas de vida aquática (CETESB, 2009).

B. Demanda Química de Oxigênio - DQO

O aumento da concentração de DQO num corpo d'água deve-se principalmente a despejos de origem industrial. O teste de DQO mede o consumo de oxigênio ocorrido em função da oxidação química da matéria orgânica, sendo o valor obtido, portanto, por uma indicação indireta do teor de matéria orgânica presente. Os valores de DQO são normalmente maiores que os da DBO_5. Como na DBO5 mede-se apenas a fração biodegradável, quanto mais este valor se aproximar da DQO, mas facilmente biodegradável será a amostra analisada (CETESB, 2009).

C. Condutividade

Representa a capacidade de condução da energia elétrica pela água a 25ºC, expressa em micro-Siemens/cm. É originada da presença de sais dissolvidos na água na forma de íons dissociados eletroliticamente. Estes íons podem ter origem antropogênica (descargas industriais, esgotos domésticos provenientes de residências e do comércio) ou geogênica (decomposição de rochas). Assim, a condutividade específica da água aumenta à medida que mais sólidos dissolvidos são adicionados. A carga de sais na água é composta por cátions (sódio, cálcio, magnésio e potássio) e ânions (cloreto, sulfato, bicarbonato, carbonato e nitrato). Altas cargas de sais na água têm seus efeitos negativos principalmente em períodos de baixa vazão.

D. Carbono Orgânico Total - COT

Neste teste, o carbono orgânico é medido diretamente, por um teste instrumental, e não indiretamente, através da determinação do oxigênio consumido, como na DBO_5 e na DQO. O teste de COT mede todo o carbono liberado na forma de CO_2. Para garantir que o carbono medido seja realmente o carbono orgânico, as formas inorgânicas de carbono (como CO_2 e HCO_3^-) devem ser removidas antes da análise ou corrigidas quando do cálculo. O teste do COT tem sido mais utilizado, até o momento, principalmente em pesquisas ou em avaliações mais aprofundadas das características do líquido, devido aos custos mais elevados do equipamento (VON SPERLING, 2005).

E. Fósforo Total

O fósforo aparece em águas naturais devido principalmente às descargas de esgotos sanitários. Nestes, os detergentes superfosfatados empregados em larga escala domesticamente constituem a principal fonte, além da própria matéria fecal, que é rica em proteínas. Alguns efluentes industriais, como os de indústrias de fertilizantes, pesticidas, químicas em geral, conservas alimentícias, abatedouros, frigoríficos e laticínios, apresentam fósforo em quantidades excessivas. As águas drenadas em áreas agrícolas e urbanas também podem provocar a presença excessiva de fósforo em águas naturais.

Assim como o nitrogênio, o fósforo constitui-se em um dos principais nutrientes para os processos biológicos, ou seja, é um dos chamados macro-nutrientes, por ser exigido também em grandes quantidades pelas células. É um nutriente essencial para o crescimento de microrganismos responsáveis pela estabilização da matéria orgânica. Usualmente os esgotos domésticos possuem um teor suficiente de fósforo, mas este pode estar deficiente em certos despejos industriais. Além disso, o fósforo é um nutriente indispensável para o crescimento de algas, mas pode, em certas condições, conduzir a fenômenos de eutrofização de lagos e represas.

F. Série do Nitrogênio (Nitrogênio Orgânico, Nitrogênio Amoniacal, Nitrito e Nitrato)

São diversas as fontes de nitrogênio nas águas naturais. Os esgotos sanitários constituem em geral a principal fonte, lançando nas águas nitrogênio orgânico devido à presença de proteínas e nitrogênio amoniacal, devido à hidrólise sofrida pela uréia na água. Alguns efluentes industriais também concorrem para as descargas de nitrogênio orgânico e amoniacal nas águas, como algumas indústrias químicas, petroquímicas, siderúrgicas,

farmacêuticas, de conservas alimentícias, matadouros, frigoríficos e curtumes. A atmosfera é outra fonte importante devido a diversos mecanismos: fixação biológica desempenhada por bactérias e algas, que incorporam o nitrogênio atmosférico em seus tecidos, contribuindo para a presença de nitrogênio orgânico nas águas, a fixação química, reação que depende da presença de luz. Concorre para a presença de amônia e nitratos nas águas, as lavagens da atmosfera poluída pelas águas pluviais concorrem para as presenças de partículas contendo nitrogênio orgânico bem como para a dissolução de amônia e nitratos. Nas áreas agrícolas, o escoamento das águas pluviais pelos solos fertilizados também contribui para a presença de diversas formas de nitrogênio. Também nas áreas urbanas, as drenagens de águas pluviais associadas às deficiências do sistema de limpeza pública, constituem fonte difusa de difícil caracterização.

Como visto, o nitrogênio pode ser encontrado nas águas nas formas de nitrogênio orgânico, amoniacal, nitrito e nitrato. As duas primeiras chamam-se formas reduzidas e as duas últimas, formas oxidadas. Pode-se associar a idade da poluição com a relação entre as formas de nitrogênio. Ou seja, se for coletada uma amostra de água de um rio poluído e as análises demonstrarem predominância das formas reduzidas significa que o foco de poluição se encontra próximo. Se prevalecer nitrito e nitrato, ao contrário, significa que as descargas de esgotos se encontram distantes. Nas zonas de autodepuração natural em rios, distinguem-se as presenças de nitrogênio orgânico na zona de degradação, amoniacal na zona de decomposição ativa, nitrito na zona de recuperação e nitrato na zona de águas limpas.

Os compostos de nitrogênio são nutrientes para processos biológicos. São tidos como macronutrientes, pois depois do carbono, o nitrogênio é o elemento exigido em maior quantidade pelas células vivas. Quando descarregados nas águas naturais conjuntamente com o fósforo e outros nutrientes presentes nos despejos, provocam o enriquecimento do meio tornando-o mais fértil e possibilitam o crescimento em maior extensão dos seres vivos que os utilizam, especialmente as algas, o que é chamado de eutrofização.

No caso do nitrogênio amoniacal, a amônia é um tóxico bastante restritivo à vida dos peixes, sendo que muitas espécies não suportam concentrações acima de 5 mg/L. Além disso, como visto anteriormente, a amônia provoca consumo de oxigênio dissolvido das águas naturais ao ser oxidada biologicamente, a chamada DBO de segundo estágio. Por estes motivos, a concentração de nitrogênio amoniacal é importante parâmetro de classificação das águas naturais e normalmente utilizado na constituição de índices de qualidade das águas.

Os nitratos podem ser considerados tóxicos, visto que podem causar uma doença chamada metahemoglobinemia infantil, que é letal para crianças (CETESB, 2009).

G. Oxigênio Dissolvido - OD

O oxigênio dissolvido é de essencial importância para os organismos aeróbios. Durante a estabilização da matéria orgânica, as bactérias fazem uso do oxigênio dissolvido nos seus processos respiratório, podendo causar uma redução da sua concentração no meio. Dependendo da magnitude deste fenômeno, podem vir a morrer diversos seres aquáticos, inclusive os peixes.

H. Potencial Hidrogeniônico - pH

Representa a concentração de íons hidrogênio H+ (em escala logarítmica), dando uma indicação sobre a condição de acidez, neutralidade ou alcalinidade da água. A faixa de pH é de 0 a 14. O valor de pH das águas limpas difere do valor neutro (pH 7) pela presença de ácido carbônico, substâncias húmicas ou pela entrada de água subterrânea com características ácidas ou alcalinas. Pode também ser influenciado pela temperatura e por sais minerais. O lançamento de efluentes nos corpos d'água e os poluentes atmosféricos (chuva ácida) também contribuem para a modificação do pH. Valores elevados de pH podem estar associados à proliferação de algas, valores elevados ou baixos podem ser indicativos da presença de efluentes industriais. Valores de pH entre 6,0 e 9,0 são considerados compatíveis a longo prazo para a sobrevivência da maioria dos organismos aquáticos. A violação destes limites por longos períodos de tempo, ou fortes oscilações de pH em curto prazo, resultam na inibição dos processos metabólicos, na redução de espécies de organismos ou no poder de autodepuração.

I. Profundidade Secchi

O disco de Secchi é utilizado para medir a transparência da coluna de água e avaliar a profundidade da zona fótica. Atualmente são utilizados discos com 20 cm de diâmetro, que podem ser inteiramente brancos, ou podem ter partes brancas e pretas alternadas.

A transparência da água medida pelo disco de Secchi varia bastante entre os ecossistemas aquáticos e, num mesmo corpo hídrico, pode variar ao longo do dia, estando na dependência do regime de circulação da massa de água, da natureza geoquímica da bacia e do regime das chuvas.

J. Sólidos

Todos os contaminantes da água, com exceção dos gases dissolvidos, contribuem para a carga de sólidos. Os sólidos podem ser classificados de acordo com o seu tamanho e

estado, sendo classificados como suspensos ou dissolvidos. O que define isto é a porosidade do filtro pelo qual a amostra irá passar. Os sólidos retidos no filtro são considerados sólidos em suspensão, ao passo que os sólidos que passam com o filtrado são considerados sólidos dissolvidos.

Os sólidos também podem ser classificados em termos da sedimentabilidade. Consideram-se como sólidos sedimentáveis aqueles que sejam capazes de sedimentar no período de 1 hora.

Valores elevados de sólidos suspensos podem indicar não apenas a contaminação orgânica recente dos rios por efluentes domésticos ou industriais, mas também um excesso de matéria sólida levada aos rios por erosão, movimentação de terra na bacia e a perda de mata ciliar.

K. Temperatura da água

A temperatura influencia todos os processos físico-químicos e biológicos da água. Também influencia a densidade e viscosidade da água alterando a sedimentação de materiais, aumentando a taxa de transferência de gases entre a água e a atmosfera, diminuindo a solubilidade dos gases na água (como no caso do oxigênio, do gás carbônico, da amônia e do nitrogênio gasoso), e aumentando a concentração de amônia livre. As origens antropogênicas deste parâmetro são o lançamento de águas de torres de resfriamento e de despejos industriais. Além disso, é importante analisar seu resultado em conjunto com outros parâmetros, como o oxigênio dissolvido (VON SPERLING, 2005).

L. Turbidez

É o grau de atenuação de intensidade que um feixe de luz sofre ao atravessar uma amostra de água devido à presença de sólidos em suspensão, tais como partículas inorgânicas (areia, silte, argila), detritos orgânicos, algas, bactérias, plâncton em geral, etc. A erosão das margens dos rios em estações chuvosas é um exemplo do fenômeno que resulta em um aumento da turbidez das águas. Os esgotos sanitários e diversos efluentes industriais também provocam elevações na turbidez das águas. A alta turbidez reduz a fotossíntese da vegetação enraizada submersa e das algas. Esse desenvolvimento reduzido de plantas pode, por sua vez, suprimir a produtividade de peixes. Logo, a turbidez pode influenciar nas comunidades aquáticas. Além disso, afeta adversamente os usos doméstico, industrial e recreacional (CETESB, 2005).

O Quadro 3.4 exibe quais foram os equipamentos utilizados para medição dos parâmetros monitorados *in situ* e a faixa de detecção para cada um deles.

QUADRO 3.4 – Parâmetros monitorados *in situ*

Parâmetro	Sensor	Marca	Faixa de Detecção
OD	Handylab OX 12/SET	SCHOTT	Escala 1: 0 a 19.99 mg/L, com resolução de 0.01 Precisão: ± 0.5% do valor medido (5° a 30°C)
pH	pH 330i/SET	WTW	Escala: -2.000 ... + 19.999, com resolução de 0.001 Precisão: ± 0.003 (15 a 35°C)
Condutividade	Handylab LF1	SCHOTT	Escala: 0.0... 199.9 µS, com resolução de 0.1 µS Precisão: ± 1% do valor medido (15°C a 35°C)
Temperatura da Água	A temperatura é lida a partir do condutivímetro e do pHmetro		pHmetro: -5.0 ... 105.0°C, com resolução de 0.1 Precisão: ± 0.1
			Condutivímetro: -5.0 ... 99.9°C, com resolução de 0.1 K Precisão: ±0.1 K
Turbidez	WQ770 Turbidimeter	Global Water	Escalas: 0 – 50 NTU ou 0 – 1000 NTU Precisão: ± 2%

O Quadro 3.5 mostra quais foram os métodos utilizados para análise de cada um dos parâmetros analisados em laboratório e suas respectivas referências de literatura, bem como, a faixa de detecção de cada teste.

QUADRO 3.5 – Parâmetros analisados em laboratório

Parâmetro	Método	Referência	Faixa de Detecção
DBO	Winkler, incubação por 5 dias a 20 ºC; determinação do OD pelo método da azida de sódio	4500 – O C e 5210 B Standard Methods (APHA, 1998)	> 2,0 mg O_2/L
DQO	Refluxo aberto	5220 B. 4b Standard Methods (APHA, 1998)	5 a 50 mg O_2/L
Nitrogênio Orgânico	Macro-Kjeldahl	4500 – N_{org} B Standard Methods (APHA, 1998)	Aplicável tanto para baixas como altas concentrações. Bastante sensível para concentrações abaixo de 5mg/L
Nitrogênio Amoniacal	Digestão seguido do método titulométrico	4500 – NH_3 B e 4500 – NH_3 C Standard Methods (APHA, 1998)	> 5mg/L
Nitrito	Colorimétrico adaptado.	4500 – NO_2^- B Standard Methods (APHA, 1998)	5 – 1000 µgNO_2^-//L
Nitrato	Redução pela coluna de cádmio/ Colorimétrico	Adaptado de 4500 - NO_3^- E Standard Methods (APHA, 1998)	0,01 a 1,0 mg NO_3^-/L
Fósforo Total	Digestão pelo ácido sulfúrico e nítrico + método colorimétrico via cloreto estanoso	4500-P B e 4500-P D Standard Methods (APHA, 1998)	0,01 a 6 mg P/L
Sólidos Sedimentáveis	Método do tubo de Inhoff	2540 F Standard Methods (APHA, 1998)	Mínimo detectável (depende da composição da amostra): 0,1 a 1,0 mL/L
Sólidos suspensos totais	Método de secagem a 103 – 105 ºC	2540 B Standard Methods (APHA, 1998)	< 200 mg de resíduo
Sólidos totais	Método da combustão a 550 ºC para fixos e voláteis	2540 E Standard Methods (APHA, 1998)	< 200 mg de resíduo
Carbono Orgânico Total	Combustão à alta temperatura, método de detecção infravermelho não dispersivo (NDIR)	TOC-V_{CPH} SHIMADZU CORPORATION, 2003	Faixa de detecção: TC < 25000 e IC < 30000 (mg/L) Limite de detecção: TC: 4 e IC : 4 (µg/L)

Os sólidos dissolvidos totais foram obtidos pela diferença dos sólidos totais e dos sólidos suspensos totais. Para a vazão foram adotados em sua maioria os valores referentes à leitura de nível relacionados à curva-chave, no entanto, para valores ausentes, optou-se por adotar os valores dos níveis obtidos da estação telemétrica.

3.5 BASE DE DADOS

A base de dados foi organizada a partir do conjunto de dados de cada um dos pontos de monitoramento, ou seja, foi formada uma única matriz contendo todas as amostras de água referentes a todos os pontos de monitoramento (*vide* Apêndice II), simbolizando a própria bacia do Alto Iguaçu. Contudo, para a realização das análises estatísticas, foram consideradas apenas as linhas que continham dados consistentes, ou seja, foram excluídas as linhas com dados de parâmetros de qualidade com falhas (ausentes) e aqueles sobre os quais recaíam algumas dúvidas. Conseqüentemente, alguns pontos de monitoramento acabaram perdendo mais linhas de dados do que outros. Deste modo, o que pode vir a ocorrer no que concerne aos resultados é que estes sejam influenciados pelos pontos de monitoramento com maior número de dados.

Uma perda considerável decorrente da adoção deste critério mais rigoroso foi que não se puderam utilizar os dados referentes ao ponto de monitoramento Olaria. Em razão de se ter decidido realizar campanhas de qualidade além de quantidade só no ano de 2008, o conjunto de dados inicial era formado somente por parâmetros monitorados *in situ* e pela vazão, não contando com os dados dos parâmetros analisados em laboratório. Além disso, no ano de 2008, houve problemas com o parâmetro da condutividade, em virtude de as amostras terem sido medidas com o condutivímetro descalibrado.

No entanto, para compensar a perda apreciável de dados, espera-se que haja ganhos na aplicação das análises e nos resultados, visto que os dados continuarão puros sem interferência alguma, senão as próprias sofridas na amostragem e nas análises laboratoriais.

A Tabela 3.1 apresenta a matriz de dados utilizada nas análises.

TABELA 3.1 – Base de dados da Bacia do Alto Iguaçu

	Nº Coleta	Data da Coleta	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SSed (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Cond (µS/cm)	T (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
P1	3	20/07/05	64	24	14	22	0,1	0,11	0,06	0,12	0,27	0,001	6,24	10,75	4,4	13,5	6,60	75	6,09	6,00
	11	14/03/06	12	6	97	20	0,1	0,65	0,16	0,04	0,35	0,022	7,80	22,31	25,1	22,1	4,88	35	6,67	2,46
	14	26/04/06	11	4	22	12	0,1	0,11	0,65	0,04	0,45	0,024	6,67	9,13	22,7	19,7	7,06	60	7,01	3,02
	17	07/06/06	15	2	21	14	0,1	0,11	0,28	0,04	0,38	0,018	5,45	9,94	18,7	16,6	7,44	70	6,70	2,40
	18	21/06/06	13	3	9	7	0,1	0,16	0,16	0,03	0,10	0,019	8,24	12,32	19,0	15,7	7,20	60	6,72	2,48
	19	19/07/06	7	3	60	8	0,1	0,33	0,38	0,03	0,05	0,035	3,62	12,10	16,6	15,1	7,52	100	6,56	2,48
P2	4	10/08/05	28	23	86	87	0,1	6,32	6,32	0,45	0,98	0,200	6,83	74,00	65,1	12,5	7,70	10	6,51	40,50
	13	10/04/06	59	40	177	34	0,2	11,43	1,47	0,06	0,35	0,123	23,25	10,75	183,6	21,3	1,40	45	7,25	8,60
	16	24/05/06	37	13	249	20	0,1	8,62	0,55	0,03	0,87	0,518	13,58	4,40	141,7	15,4	1,86	45	7,30	8,60
	17	07/06/06	39	18	149	32	0,1	8,85	1,34	0,05	0,19	1,820	18,15	12,83	163,1	17,2	2,32	30	7,00	7,76
	18	21/06/06	64	25	143	18	0,4	1,7	1,26	0,08	0,08	1,500	24,59	14,87	171,1	16,5	2,02	15	7,28	7,55
P3	3	20/07/05	42	92	181	27	0,1	7,43	0,22	0,44	0,20	0,534	12,22	12,07	126,1	13,3	1,40	50	6,66	12,00
	7	19/10/05	25	10	81	48	0,4	1,15	1,50	0,34	0,19	0,200	9,74	20,60	72,1	19,1	3,48	45	6,59	24,59
	11	14/03/06	28	7	120	29	0,1	6,79	4,68	0,05	0,92	0,223	7,70	18,55	114,3	23,4	3,22	20	6,80	30,04
	12	03/04/06	30	8	175	20	0,1	5,77	1,02	0,02	1,10	0,888	9,78	14,96	73,4	21,1	1,80	40	7,03	20,75
	13	10/04/06	24	10	133	31	0,3	7,01	1,58	0,04	0,34	0,105	12,53	11,47	125,6	21,4	1,34	50	7,16	9,06
	14	26/04/06	35	26	119	12	0,1	10,06	1,90	0,07	0,60	1,300	14,96	10,88	139,2	21,2	0,92	40	7,25	8,54
	16	24/05/06	32	35	101	34	0,5	0,88	0,33	0,03	0,12	0,467	21,22	3,20	120,0	14,6	1,92	35	7,20	8,89
	17	07/06/06	36	26	147	24	0,1	8,01	1,40	0,06	0,42	1,870	20,83	16,91	163,9	18,0	0,76	40	7,00	9,06
	18	21/06/06	35	22	311	35	0,2	9,54	2,14	0,07	0,09	1,440	45,63	17,79	175,3	16,9	0,56	20	7,33	8,54
P 4	7	19/10/05	10	8	41	36	0,1	2,31	1,38	0,54	0,19	0,168	6,65	22,38	80,4	19,1	1,82	25	6,62	74,22
	12	03/04/06	13	11	167	13	0,1	6,05	1,70	0,08	0,44	0,561	11,49	11,34	75,9	20,7	1,76	15	7,06	59,97
	17	07/06/06	31	9	129	51	0,2	7,56	1,29	0,05	0,95	1,550	21,37	16,87	166,9	18,2	0,66	30	7,10	17,03
P5	7	19/10/05	19	10	72	41	0,3	2,02	1,10	0,64	0,23	0,141	7,26	25,72	73,4	19,7	2,16	50	6,59	84,08
	13	10/04/06	18	7	186	41	0,1	5,09	0,85	0,08	0,49	0,054	8,48	21,55	123,3	21,1	1,60	45	7,23	26,70
	14	26/04/06	13	7	136	11	0,1	4,78	1,58	0,06	0,55	0,820	9,25	16,19	129,9	20,9	2,14	60	7,24	19,30
	16	24/05/06	29	6	188	47	0,5	0,6	0,05	0,05	0,15	0,232	10,77	13,77	108,1	14,0	1,00	35	7,20	27,00
	17	07/06/06	23	9	94	35	0,1	6,61	1,18	0,05	0,20	1,230	13,30	12,19	150,8	18,3	0,98	60	7,00	19,98
	18	21/06/06	23	13	155	15	0,1	8,34	1,98	0,14	0,08	1,010	13,02	13,43	162,1	17,1	0,96	60	7,20	14,60
P6	5	19/10/05	9	7	81	28	0,1	1,56	1,21	0,77	0,25	0,071	5,71	23,16	62,8	20,0	2,46	40	6,60	103,38
	12	26/04/06	13	7	135	16	0,1	5,16	0,54	0,13	0,69	0,555	7,17	12,96	113,3	20,8	4,06	40	7,40	20,07
	14	24/05/06	26	9	241	38	0,1	6,09	0,55	0,03	0,62	0,306	22,65	6,97	99,9	14,7	1,16	40	7,30	30,44
	15	07/06/06	17	10	53	21	0,1	5,66	1,01	0,06	0,08	0,958	9,12	12,83	135,0	17,8	2,60	50	7,10	8,11
	16	21/06/06	18	6	255	5	0,1	10,15	1,04	0,09	0,07	0,910	10,40	11,00	149,9	17,6	2,56	60	7,47	16,47

3.6 APLICAÇÃO DOS MÉTODOS NA BACIA DO ALTO IGUAÇU

Os métodos foram empregados de acordo com a ordem dos objetivos específicos apresentados no item 1.2.2. Assim, inicialmente realizou-se a Análise de Componentes Principais dos dados de monitoramento de qualidade de água da bacia do Alto Iguaçu, objetivando identificar os parâmetros de qualidade de água mais representativos na bacia. Para tanto foi utilizado o *software* MATLAB 5.3, com o emprego da função programada **comp2**, descrita no Anexo I. Através desta função, as componentes principais foram extraídas da matriz de correlação, eliminando possíveis influências em virtude das diferentes ordens de magnitude dos parâmetros de qualidade de água. Para estimação do número de componentes principais utilizou-se o critério de Kaiser. Sob este critério, foram retidas as componentes com autovalores maiores que 1. A Análise de Componentes Principais também foi utilizada para o desenvolvimento do terceiro objetivo específico que consta no item 1.2.2, o qual trata da avaliação dos pontos de monitoramento da bacia.

Ainda com o mesmo intuito, o segundo método estatístico multivariado aplicado foi a Análise Fatorial. Antes de se iniciar a análise, no entanto, foram realizados alguns testes. Estes testes também foram efetuados no *software* MATLAB versão 5.3. A normalidade multivariada dos dados foi testada através da função programada **normult**, descrita no Anexo II, como exposto no item 2.1.2. Caso a normalidade fosse comprovada, poderia se utilizar o método de máxima verossimilhança para extração dos fatores, caso contrário poderia se utilizar o método das componentes principais que não requer a normalidade dos dados. Para avaliar se a estrutura dos dados era adequada à análise fatorial, foram realizados os testes de esfericidade de Bartlett – demonstrado no item 2.3.1 - e da medida de adequacidade da amostra de Kaiser-Meyer-Olkin (KMO) – demonstrada no item 2.3.2. Para aplicação destes testes, utilizou-se a função programada **KMO**, descrita no Anexo III. Após a realização desses testes, a Análise Fatorial foi realizada no *software* STATISTICA versão 6.0. A escolha deste software deveu-se à possibilidade de realizar a rotação varimax, que é a rotação dos fatores (item 2.3.8), a qual tem por objetivo obter pesos altos para cada variável em um único fator e pesos baixos ou moderados nos demais fatores.

Para o agrupamento das amostras de água coletadas no rio, empregou-se a análise de agrupamentos. Assim, previamente calculou-se a correlação cofenética pela função **cophenet** programada no MATLAB (Anexo IV). Esta correlação mostrou a melhor "combinação" entre distâncias e ligações (itens 2.4.1 a 2.4.3). Esta correlação é na verdade a correlação entre a ligação dos objetos no agrupamento com as distâncias entre objetos no vetor de distâncias. Deste modo, foram escolhidas a distância e a ligação que apresentaram correlação mais próxima de 1. Após o cálculo da correlação e escolhidos os tipos de distância e ligação a serem utilizados, os dados foram analisados pelo método estatístico

Cluster Analysis (ou Análise de Agrupamentos) do *software* STATISTICA para determinação dos agrupamentos.

3.7 ESTRATÉGIAS DE AVALIAÇÃO

Para a realização da análise multivariada dos dados de monitoramento de qualidade de água foram adotadas duas estratégias de avaliação para análise dos dados. Na primeira estratégia de avaliação, as variáveis analisadas foram os parâmetros de qualidade de água, e, na segunda, foram os pontos de monitoramento. Os principais objetivos destas duas estratégias foram identificar a relevância dos parâmetros de qualidade de água e dos pontos de monitoramento na avaliação da qualidade da água da bacia, e, evidenciar possíveis relações existentes no âmbito de cada grupo.

A identificação dos parâmetros de qualidade de água e das estações de monitoramento mais representativas pode apontar para uma nova estratégia de monitoramento, na qual poderia se optar por monitorar os parâmetros considerados menos significantes com uma freqüência menor do que a de costume e o mesmo para os pontos de monitoramento. Isto resultaria em um menor tempo gasto em campanhas de monitoramento e em análises laboratoriais, e, na redução de custos. Em casos mais extremos, poderia se optar até mesmo pelo descarte de parâmetros de qualidade de água e desativação de estações de monitoramento, no entanto, para se chegar a tal sentença, seria necessário estender o assunto com estudos mais aprofundados e completos.

Para a primeira avaliação - **Análise I** - foi realizada uma análise considerando-se a bacia como um todo, sendo os dados de qualidade de água agrupados em uma mesma matriz conforme exposto na Tabela 3.1. Deste modo, esta análise compreendeu alterações espaciais e temporais simultaneamente, utilizando-se para tanto das técnicas estatísticas de componentes principais e fatorial.

Na **Análise I**, também foi avaliada a semelhança entre amostras de água, visando o agrupamento das amostras da água do rio com base nas suas composições, buscando identificar quais amostras foram coletadas quando o rio apresentava melhores condições de qualidade, em que momento isto ocorreu e em que local de monitoramento. Para tanto, a técnica utilizada foi a análise de agrupamentos através do método de agrupamento hierárquico. Para escolha da distância e do tipo de ligação a serem utilizados para formação dos agrupamentos, foi calculada a correlação cofenética através da função **cophenet**[11] do *software* MATLAB.

A **Análise II** contou com a mesma base de dados, porém como mencionado, neste caso as variáveis foram os pontos de monitoramento. Nas colunas da matriz, foram

[11] O algoritmo desta função é apresentado no Anexo IV

dispostos os pontos de monitoramento e, nas linhas, os próprios parâmetros de qualidade de água. Estes foram representados pela mediana dos dados coletados em diferentes campanhas (Tabela 3.2), reproduzindo-se então o modelo da base de dados do trabalho realizado por Ouyang (2005), exposto no item 2.5.2.

No presente estudo, no entanto, esta análise pode ter sofrido influências, visto que para o P4, por exemplo, a mediana resultou de três dados, enquanto que para o P3, a mediana foi calculada para nove dados.

TABELA 3.2 – Base de dados para a Análise II

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
DQO (mg/L)	13	39	32	13	21	17
DBO_5 (mg/L)	4	23	22	9	8	7
Sólidos Dissolvidos Totais (mg/L)	22	149	133	129	145	135
Sólidos Suspensos Totais (mg/L)	13	32	29	36	38	21
Sólidos Sedimentáveis (mL/L)	0,10	0,10	0,10	0,10	0,10	0,10
N-Amoniacal (mg/L)	0,14	8,62	7,01	6,05	4,94	5,66
N-Orgânico (mg/L)	0,22	1,34	1,50	1,38	1,14	1,01
Nitrito (mg/L)	0,04	0,06	0,06	0,08	0,07	0,09
Nitrato (mg/L)	0,31	0,35	0,34	0,44	0,22	0,25
Fósforo (mg/L)	0,02	0,52	0,53	0,56	0,53	0,56
COT (mg/L)	6,46	18,15	12,53	11,49	10,01	9,12
Turbidez (NTU)	11,43	12,83	14,96	16,87	14,98	12,83
Condutividade (µS/cm)	18,85	163,10	125,60	80,40	126,60	113,30
Temperatura (°C)	16,15	16,50	19,10	19,10	19,00	17,80
OD (mg/L)	7,13	2,02	1,40	1,76	1,30	2,56
Profundidade Secchi (cm)	65,00	30,00	40,00	25,00	55,00	40,00
pH	6,69	7,25	7,03	7,06	7,20	7,30
Vazão (m^3/s)	2,48	8,60	9,06	59,97	23,34	20,87

3.8 SÍNTESE DO CAPÍTULO

Neste capítulo, foi apresentada a bacia do Alto Iguaçu, incluindo seus aspectos físico, climático e demográfico, e, a localização dos pontos de monitoramento de qualidade de água em sua extensão. Também foram descritos os parâmetros de qualidade de água utilizados no monitoramento da bacia, os equipamentos utilizados *in situ* e as análises realizadas em laboratório. Discutiram-se ainda os métodos a serem empregados para aplicação das técnicas multivariadas escolhidas e as estratégias de avaliação do conjunto de dados.

CAPÍTULO IV

4. RESULTADOS

Neste capítulo são apresentados os resultados das aplicações dos métodos estatísticos multivariados conforme discutido nos itens 3.6 e 3.7. Foram consideradas duas abordagens: (1) dados dos distintos pontos de monitoramento da bacia do Alto Iguaçu tratados globalmente (Tabela 3.1), denominada Análise Global da Bacia do Alto Iguaçu e (2) mediana dos dados dos parâmetros de qualidade de água individualizados por ponto de monitoramento (Tabela 3.2), denominada Análise dos Pontos de Monitoramento. Ou seja, na primeira análise as variáveis foram os próprios parâmetros de qualidade de água e na segunda, as variáveis foram os pontos de monitoramento.

Objetivou-se com estas análises identificar a relevância dos parâmetros de qualidade e dos pontos de monitoramento na avaliação da qualidade da água da Bacia do Alto Iguaçu, bem como as relações existentes entre os parâmetros de qualidade de água e entre os pontos de monitoramento .

4.1 ANÁLISE GLOBAL DA BACIA DO ALTO IGUAÇU

Nesta análise foi avaliada uma matriz de ordem 34 X 18 (linhas X colunas), denominada Amostra I, apresentada na Tabela 3.1. Como variáveis, consideraram-se os 18 parâmetros de qualidade de água e como observações, as 34 campanhas de monitoramento realizadas nos pontos de monitoramento P1, P2, P3, P4, P5 e P6 da bacia. As datas e os pontos de monitoramento nos quais as observações foram feitas encontram-se no Quadro 4.1. Os parâmetros avaliados foram DBO_5, DQO, sólidos dissolvidos totais, sólidos suspensos totais, sólidos sedimentáveis, nitrogênio amoniacal, nitrogênio orgânico, nitrito, nitrato, fósforo, COT, turbidez, condutividade, temperatura da água, OD, profundidade Secchi, pH e vazão. É importante ressaltar que esta análise pode ser dita como mais rigorosa, visto que se optou por considerar apenas as linhas de observações que não apresentavam falhas, ou seja, dados faltantes e duvidosos. Um exemplo de dados duvidosos seriam os valores de condutividade medidos sem que o condutivímetro apresentasse uma calibração confiável. Assim, a Amostra I é um conjunto de todos os dados coletados na bacia, considerando todos os pontos, excluindo-se, no entanto, dados que poderiam vir a comprometer as análises. Além disso, vale enfatizar que mesmo que só um parâmetro apresentasse falha, descartou-se a coleta inteira.

Nesta análise, serão realizadas as análises de componentes principais, fatorial e agrupamentos.

QUADRO 4.1 - Observações

Observações	Número da Coleta	Ponto de Monitoramento	Data da Coleta
1	3	P1	20/07/2005
2	11	P1	14/03/2006
3	14	P1	26/04/2006
4	17	P1	07/06/2006
5	18	P1	21/06/2006
6	19	P1	19/07/2006
7	4	P2	10/08/2005
8	13	P2	10/04/2006
9	16	P2	24/05/2006
10	17	P2	07/06/2006
11	18	P2	21/06/2006
12	3	P3	20/07/2005
13	7	P3	19/10/2005
14	11	P3	14/03/2006
15	12	P3	03/04/2006
16	13	P3	10/04/2006
17	14	P3	26/04/2006
18	16	P3	24/05/2006
19	17	P3	07/06/2006
20	18	P3	21/06/2006
21	7	P4	19/10/2005
22	12	P4	03/04/2006
23	17	P4	07/06/2006
24	7	P5	19/10/2005
25	13	P5	10/04/2006
26	14	P5	26/04/2006
27	16	P5	24/05/2006
28	17	P5	07/06/2006
29	18	P5	21/06/2006
30	5	P6	19/10/2005
31	12	P6	26/04/2006
32	14	P6	24/05/2006
33	15	P6	07/06/2006
34	16	P6	21/06/2006

4.1.1 Estatística descritiva das 18 variáveis

A Tabela 4.1 apresenta a média, a variância, o desvio padrão e os coeficientes de variação de cada uma das variáveis originais referentes à Amostra I (Tabela 3.1), as variáveis foram dispostas em ordem decrescente do valor do coeficiente de variação.

TABELA 4.1- Estatística descritiva das 18 variáveis

Variável Original	Média	Desvio Padrão	Variância	Coeficiente de Variação
Nitrito (mg/L)	0,14	0,19	0,04	1,36
Vazão (m^3/s)	21,90	24,23	587,29	1,11
DBO (mg/L)	15,17	16,55	273,84	1,09
Fósforo (mg/L)	0,58	0,58	0,34	0,99
Nitrogênio Orgânico (mg/L)	1,26	1,24	1,54	0,98
OD (mg/L)	2,86	2,27	5,16	0,79
Nitrato (mg/L)	0,38	0,30	0,09	0,78
Sólidos Sedimentáveis (mL/L)	0,16	0,12	0,01	0,75
Nitrogênio Amoniacal (mg/L)	4,91	3,55	12,58	0,72
Turbidez (NTU)	15,89	11,50	132,35	0,72
COT (mg/L)	12,81	8,20	67,22	0,64
Sólidos Suspensos Totais (mg/L)	27,41	16,29	265,34	0,59
DQO (mg/L)	26,34	14,93	222,78	0,57
Sólidos Dissolvidos Totais (mg/L)	127,25	72,86	5.308,69	0,57
Condutividade (µS/cm)	105,08	53,10	2.819,30	0,51
Profundidade Secchi (cm)	43,97	18,70	349,67	0,43
Temperatura (ºC)	18,08	2,87	8,25	0,16
pH	6,98	0,32	0,11	0,05

Como se torna difícil a comparação do desvio padrão entre variáveis de diferentes grandezas, optou-se por calcular o coeficiente de variação, que é igual ao desvio padrão dividido pela média. Assim, é possível comparar a variação de conjuntos de observações que diferem na média ou são medidos em unidades de medição diferentes, e, classificar o grau de dispersão das variáveis. Neste trabalho, considerou-se o seguinte critério para avaliação do grau de dispersão:

QUADRO 4.2 – Critério de avaliação do grau de dispersão

Valor do Coeficiente de Variação	Grau de Dispersão
< 0,50	Baixo
0,50 - 1,00	Médio
> 1,00	Alto

Deste modo, pode-se dizer que o nitrito, a vazão e a DBO apresentaram um alto grau de dispersão. Este alto grau de dispersão para o nitrito pode ser explicado em razão de este parâmetro não ser estável conforme resultados de monitoramento. Para a vazão, esta variação ocorre em virtude de as campanhas terem sido realizadas tanto em épocas de cheia como de estiagem. Uma conseqüência, em especial no caso da DBO, para explicação desta variação é que as coletas foram realizadas desde pontos em áreas de manancial até pontos em regiões notadamente marcadas pela poluição.

As variáveis que apresentaram grau de dispersão médio são diretamente afetadas pela variação da vazão que pode resultar tanto nos seus incrementos como nas suas diluições.

Por outro lado, o pH, a temperatura da água, a profundidade Secchi e a condutividade – que não são variáveis que se modificam com a vazão - apresentaram um baixo grau de dispersão. Os dados de pH e de temperatura variaram muito pouco durante o período monitorado ao longo das estações de monitoramento,mostrando-se estáveis ao longo das coletas. A condutividade e a profundidade Secchi que são relacionadas com os sólidos apresentaram variação um pouco maior, justamente em razão desta relação de dependência com outras variáveis, que no caso tiveram grau de dispersão médio.

No geral, pode-se afirmar que a maioria dos parâmetros de qualidade de água apresentou uma dispersão considerável, o que pode ser explicado pela própria variabilidade natural associada aos seus dados que sofrem influências temporais e espaciais.

Uma decorrência importante das variâncias das variáveis é perceber que caso fosse utilizada uma rotina de componentes principais que extraísse os autovalores e respectivos autovetores da matriz de covariância, os resultados da análise seriam comprometidos pela influência exercida pelas variáveis de maior variância. Assim, muitas vezes opta-se por calcular as componentes principais a partir da matriz de correlação dos dados originais, que elimina as influências exercidas pelas diferentes magnitudes das variáveis consideradas.

4.1.2 Matriz de Correlação das 18 variáveis

A Tabela 4.2, apresentada a seguir, exibe as correlações existentes entre as 18 variáveis. Os valores em vermelho são aqueles superiores ou iguais a $|0{,}5|$.

Neste trabalho, em virtude da natural variabilidade dos parâmetros de qualidade de água, o que muitas vezes reflete em baixas correlações entre os parâmetros, considerou-se que uma correlação de "$|0{,}5|$" já seria razoável para se afirmar que os parâmetros de qualidade de água estão relacionados entre si.

Vega *et al*. (1998) sugerem que as correlações sejam interpretadas com cautela quando combinadas diferentes estações de monitoramento – que é do que se trata a presente análise - visto que são afetadas tanto espacial como temporalmente.

TABELA 4.2 – Matriz de correlação das 18 variáveis

	DQO	DBO_5	SDT	SST	SSed	N-A	N-Org	NO_2^-	NO_3^-	Fósforo	COT	Turbidez	Cond	T	OD	Secchi	pH	Q
DQO	**1,00**	0,57	0,26	0,17	0,29	0,30	0,05	-0,17	0,03	0,34	0,51	-0,10	0,41	-0,29	-0,25	-0,22	0,05	-0,34
DBO_5	0,57	**1,00**	0,22	0,13	0,09	0,30	-0,002	0,19	-0,12	0,16	0,31	-0,02	0,31	-0,34	-0,24	-0,13	-0,06	-0,16
SDT	0,26	0,22	**1,00**	0,06	0,06	0,68	0,06	-0,24	0,15	0,39	0,63	-0,16	0,66	-0,01	-0,65	-0,34	0,69	-0,09
SST	0,17	0,13	0,06	**1,00**	0,32	0,07	0,52	0,37	0,27	-0,07	0,15	0,70	0,12	-0,24	-0,11	-0,49	-0,20	0,32
SSed	0,29	0,09	0,06	0,32	**1,00**	-0,30	-0,13	0,02	-0,30	-0,03	0,28	-0,09	0,17	-0,17	-0,25	-0,23	0,15	0,001
N-A	0,30	0,30	0,68	0,07	-0,30	**1,00**	0,37	-0,19	0,28	0,56	0,48	-0,02	0,78	0,18	-0,59	-0,27	0,55	-0,14
N-Org	0,05	-0,002	0,06	0,52	-0,13	0,37	**1,00**	0,21	0,42	0,11	0,05	0,74	0,19	0,12	0,04	-0,51	-0,08	0,23
NO_2^-	-0,17	0,19	-0,24	0,37	0,02	-0,19	0,21	**1,00**	-0,11	-0,27	-0,28	0,50	-0,20	-0,04	0,01	-0,16	-0,51	0,81
NO_3^-	0,03	-0,12	0,15	0,27	-0,30	0,28	0,42	-0,11	**1,00**	0,005	-0,11	0,29	-0,003	0,26	0,01	-0,33	0,09	0,05
Fósforo	0,34	0,16	0,39	-0,07	-0,03	0,56	0,11	-0,27	0,005	**1,00**	0,59	-0,13	0,72	-0,01	-0,54	-0,29	0,47	-0,25
COT	0,51	0,31	0,63	0,15	0,28	0,48	0,05	-0,28	-0,11	0,59	**1,00**	-0,18	0,66	-0,13	-0,54	-0,41	0,51	-0,26
Turbidez	-0,10	-0,02	-0,16	0,70	-0,09	-0,02	0,74	0,50	0,29	-0,13	-0,18	**1,00**	-0,16	-0,10	0,29	-0,39	-0,38	0,37
Cond	0,41	0,31	0,66	0,12	0,17	0,78	0,19	-0,20	-0,003	0,72	0,66	-0,16	**1,00**	0,11	-0,80	-0,37	0,72	-0,15
T	-0,29	-0,34	-0,01	-0,24	-0,17	0,18	0,12	-0,04	0,26	-0,01	-0,13	-0,10	0,11	**1,00**	-0,25	-0,16	0,21	0,20
OD	-0,25	-0,24	-0,65	-0,11	-0,25	-0,59	0,04	0,01	0,01	-0,54	-0,54	0,29	-0,80	-0,25	**1,00**	0,40	-0,59	-0,18
Secchi	-0,22	-0,13	-0,34	-0,49	-0,23	-0,27	-0,51	-0,16	-0,33	-0,29	-0,41	-0,39	-0,37	-0,16	0,40	**1,00**	-0,23	-0,32
pH	0,05	-0,06	0,69	-0,20	0,15	0,55	-0,08	-0,51	0,09	0,47	0,51	-0,38	0,72	0,21	-0,59	-0,23	**1,00**	-0,27
Q	-0,34	-0,16	-0,09	0,32	0,001	-0,14	0,23	0,81	0,05	-0,25	-0,26	0,37	-0,15	0,20	-0,18	-0,32	-0,27	**1,00**

Observa-se que as variáveis sólidos sedimentáveis (SSed), nitrato (NO_3^-) e temperatura da água (T) não obtiveram correlação maior ou igual a $|0,5|$ com nenhuma outra variável. A Tabela 4.3 exibe um resumo das correlações destacadas na Tabela 4.2.

TABELA 4.3 – Resumo das correlações

Variável 1	Variável 2	Correlação
DQO	DBO_5	0,57
	COT	0,51
SDT	N-A	0,68
	COT	0,63
	Condutividade	0,66
	OD	-0,65
	pH	0,69
SST	N-Org	0,52
	Turbidez	0,70
N-A	Fósforo	0,56
	Condutividade	0,78
	OD	-0,59
	pH	0,55
N-Org	Turbidez	0,74
	Secchi	-0,51
NO_2^-	Turbidez	0,50
	pH	-0,51
	Q	0,81
Fósforo	COT	0,59
	Condutividade	0,72
	OD	-0,54
COT	Condutividade	0,66
	OD	-0,54
	pH	0,51
Condutividade	OD	0,80
	pH	-0,72
OD	pH	-0,59

4.1.3 Análise de Componentes Principais

A análise de componentes principais foi realizada utilizando-se o *software* MATLAB versão 5.3, através da função programada **comp2**[12]. Nesta função, os autovalores e respectivos autovetores são obtidos diretamente da matriz de correlação, evitando possíveis incoerências devido à diferença de unidades e escalas dos valores medidos. Os resultados obtidos para a análise de componentes principais são apresentados na seqüência.

a) Estimação do Número de Componentes Principais

Para estimação do número de componentes principais, primeiramente obteve-se a matriz de correlação da Amostra I, que é a própria Tabela 4.2. Foram calculados então os autovetores e autovalores da matriz de correlação. Os autovalores, em ordem decrescente, definem a importância das componentes principais. Os autovalores correspondem à variância explicada por cada uma das componentes principais. Os valores são apresentados na Tabela 4.4:

TABELA 4.4 – Autovalores e variância total

Componente Principal	*Autovalor*	*Variância Explicada (%)*	*Variância Explicada Acumulada (%)*
1	5,40	30,02	30,02
2	3,46	19,25	49,27
3	2,26	12,53	61,81
4	1,70	9,47	71,28
5	1,26	7,00	78,27
6	0,83	4,59	82,86
7	0,75	4,19	87,05
8	0,53	2,95	90,00
9	0,47	2,60	92,60
10	0,35	1,95	94,55
11	0,29	1,60	96,14
12	0,19	1,04	97,18
13	0,17	0,97	98,15
14	0,16	0,87	99,03
15	0,07	0,40	99,43
16	0,05	0,28	99,70
17	0,04	0,20	99,90
18	0,02	0,10	100,00

[12] Código fonte da função comp2 está descrito no Anexo I

Para a escolha do número de componentes, adotando-se o critério de Kaiser (KAISER, 1958), no qual o número de autovalores é igual ao número de autovalores maiores que 1 (item 2.2.4), obtiveram-se 5 componentes principais e a porcentagem da variância explicada por elas foi igual a 78,27%, conforme mostra a Tabela 4.4. Pelo método *Scree-Plot* (CATTELL, 1966), resultariam 10 componentes principais responsáveis por uma variância explicada acumulada igual a 94,55%. Neste trabalho, no entanto, optou-se por utilizar o critério de Kaiser e considerou-se suficiente a variância explicada pelas 5 primeiras componentes principais. Isto porque se percebe na Tabela 4.4 que a partir da sexta componente principal, a variância explicada é praticamente marginal, quando comparada com as variâncias das componentes de 1 a 5. A Figura 4.1 mostra a comparação entre os dois critérios e a variância explicada acumulada pelas componentes.

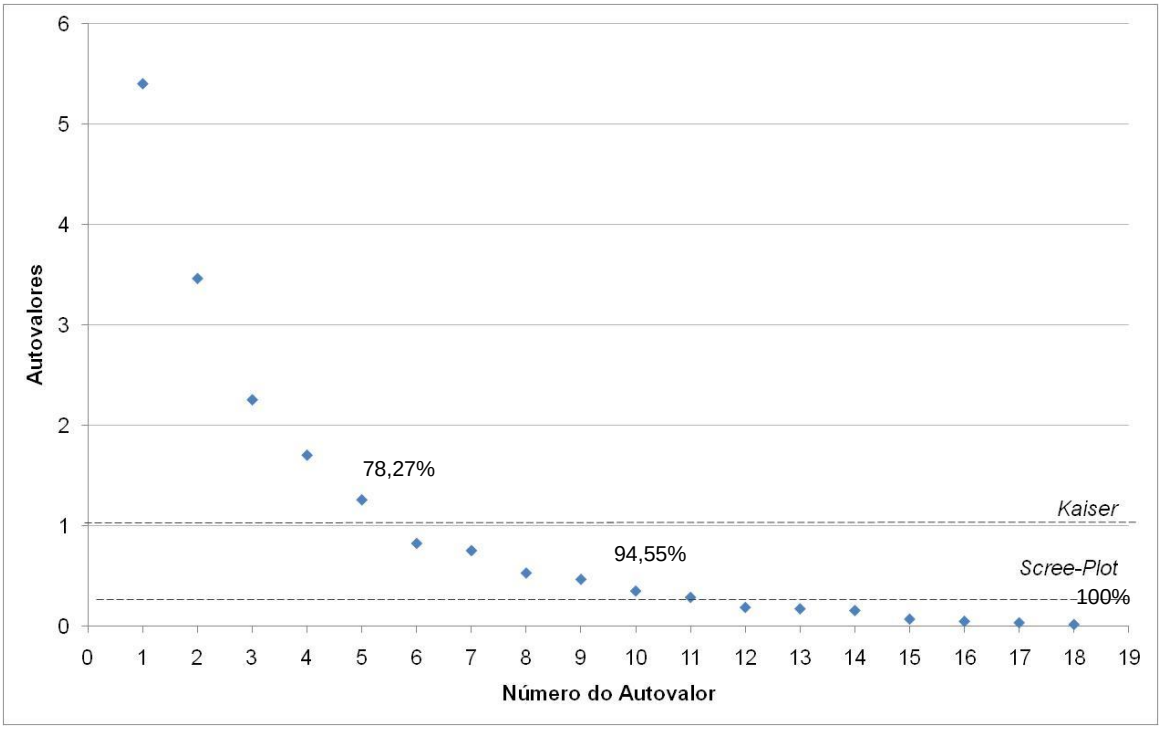


FIGURA 4.1 - Autovalores: *Scree Plot* X Kaiser

b) Componentes Principais da Amostra I

Os autovetores da matriz de correlação têm sua importância definida pelos autovalores. Assim, a 1ª componente principal refere-se ao maior autovalor e a última componente ao menor autovalor. Os valores que constituem os autovetores representam os pesos (ou carregamentos ou *loadings*) das variáveis originais nas componentes principais, ou ainda, os coeficientes das variáveis na combinação linear que é a própria componente principal.

Os pesos das variáveis nas componentes principais representam uma indicação da importância de cada um dos parâmetros nas componentes, o que é confirmado pelo cálculo das correlações entre variáveis e componentes. Na Figura 4.2, são apresentados os pesos (em azul) e as correlações (em vermelho) de cada uma das variáveis originais - parâmetros de qualidade de água - nas 5 componentes principais.

Variáveis que apresentaram correlações maiores ou iguais a 0,7 – em valores absolutos – foram consideradas importantes para a definição das componentes principais.

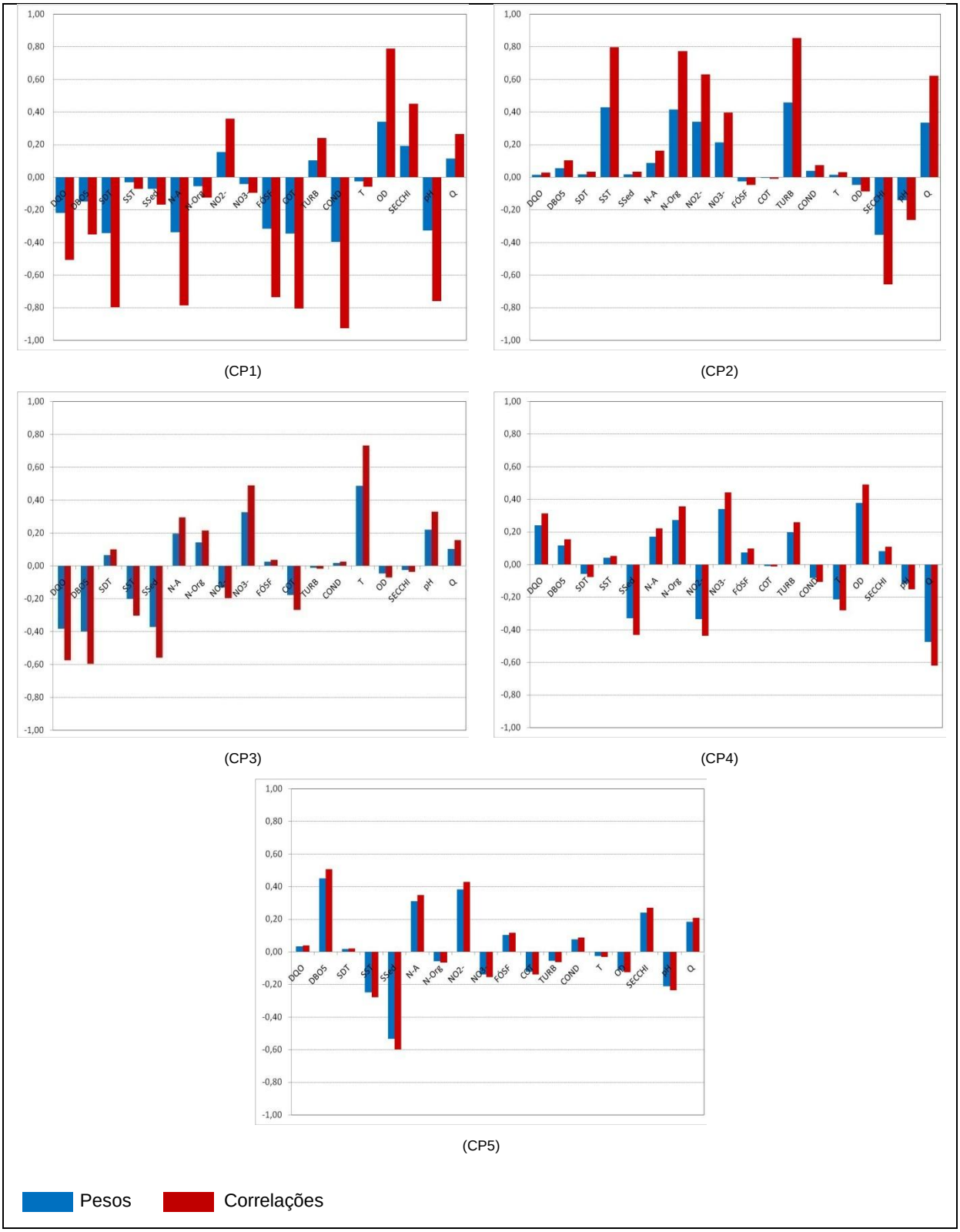


FIGURA 4.2- Pesos e correlações das variáveis

As variáveis que mais se destacaram em cada uma das 5 componentes principais constam na Tabela 4.5. A tabela também apresenta a variância total explicada por cada uma das 5 componentes principais retidas, o que auxilia a avaliar a importância de cada uma das componentes na explicação do conjunto original de dados, a Amostra I.

TABELA 4.5 – Variáveis com maior peso na definição das componentes principais

	Componente Principal	*Variância (%)*	*Variáveis com correlação* ≥ | *0,7*|
CPs com Alta Variância	1	30,02	OD (+), SDT (-), N-A (-), Fósforo (-), COT (-), Condutividade (-), pH (-)
	2	19,25	SST (+), N-Org (+), Turbidez (+)
CPs com Baixa Variância	3	12,53	Temperatura (+)
	4	9,47	Nenhuma variável
	5	7,00	Nenhuma variável

Na primeira componente principal, que representa cerca de 30% da variância explicada, destacaram-se as variáveis SDT, N-A, Fósforo, COT e pH com pesos altos e negativos e OD com peso alto e positivo (Figura 4.2, CP1). Esta componente expressa os aspectos de degradação da matéria orgânica (COT, N-A), resultante de esgotos, e sua interação com a dinâmica de transportes de sólidos.

A diferença entre os sinais dos pesos e das correlações das variáveis indica justamente o contraste entre elas: enquanto que para o OD, quanto maior for a concentração melhor será a qualidade da água, para o COT, por exemplo, será exatamente o oposto. Nota-se que a condutividade e os SDT apresentam o mesmo sinal, ilustrando a relação existente entre eles, na qual quanto maior for a entrada de sólidos dissolvidos na bacia, maior será o valor da condutividade elétrica.

Na segunda componente principal, que explica uma variância total de 19,25% da amostra, os parâmetros de qualidade de água que apresentaram correlações altas e positivas foram sólidos suspensos totais, nitrogênio orgânico e turbidez. Pode-se afirmar que esta componente destaca a importância do nitrogênio na poluição orgânica, no entanto, a vazão não apresenta sinal oposto às concentrações de nitrogênio, o que pode indicar que não auxilia na diluição da carga poluidora. Pode indicar justamente o contrário, que se trata de uma vazão relacionada à poluição difusa da bacia, o que compromete ainda mais o estado de qualidade de água da bacia.

Além disso, um resultado interessante para as componentes principais 1 e 2 é que nelas ficaram relacionadas as formar reduzidas de nitrogênio (amoniacal e orgânica), o que aponta para uma poluição mais recente, ou seja, significa que o foco de poluição se encontra próximo.

Caso se aceitasse uma correlação absoluta de 0,6, seriam agregados ao conjunto ainda o nitrito e a vazão, com valores positivos, e a profundidade Secchi, com valor negativo (Figura 4.2, CP2).

Os sólidos suspensos e a turbidez avançam no mesmo sentido ilustrando a relação existente entre eles e indicando a poluição da parte estética do rio.

Na componente principal 3 (Figura 4.2, CP3), a única variável com correlação absoluta maior que 0,7 foi a temperatura da água. A DQO e a DBO_5 foram as variáveis que apareceram em segundo lugar de importância, caso fossem aceitas correlações inferiores a $\left|0{,}7\right|$. No entanto, quando uma CP substitui apenas 1 variável, pode-se optar por trabalhar diretamente com a variável original.

É interessante salientar, neste momento, que nem sempre foi possível extrair informações de boa qualidade de todas as componentes principais, sendo imprudente produzir artificialmente interpretações que não acrescentem novas informações. Além disso, as últimas componentes exibiram variâncias baixas, que não deram margem a interpretações confiáveis.

Para as componentes 4 e 5 (Figura 4.2, CP4 e CP5, respectivamente), nenhuma variável atingiu correlação igual a $\left|0{,}7\right|$. Para a componente 4, a vazão apresentou a maior correlação negativa e o OD a maior correlação positiva. O que indicou novamente que a vazão poderia não estar auxiliando na diluição ou mesmo na reaeração, mas contribuindo negativamente para a qualidade da água, pelo menos em um primeiro instante. O que pode ser explicado em razão de a chuva ter carreado poluentes para o rio e ter se iniciado a primeira fase da decomposição, onde o consumo de oxigênio é maior. Assim, mesmo que tenha havido aumento da vazão no início, não houve efeito de diluição.

A quinta componente principal, responsável por 7% da variância total explicada, teve entre os parâmetros que mais se destacaram a DBO_5, e NO_2^- com pesos altos e

positivos e sólidos sedimentáveis com peso alto e negativo. No entanto, nenhuma variável atinge correlação igual ou superior a │0,7│ (Figura 4.2, CP5).

Para sintetizar as informações obtidas para as componentes principais 1 e 2, utilizou-se a representação gráfica dos pesos das variáveis originais nestas componentes. Observa-se na Figura 4.3, a formação dos grupos de variáveis já mencionados. Assim, tem-se na CP1 as variáveis SDT, DQO, COT, pH, condutividade, nitrogênio amoniacal e fósforo no lado esquerdo do gráfico e OD no lado direito. As variáveis próximas ao zero do gráfico são aquelas que não têm pesos significantes em nenhuma das 2 componentes. Na CP2, as variáveis agrupadas foram SST, nitrogênio orgânico, nitrito, turbidez e vazão na parte superior do gráfico e a profundidade Secchi na parte inferior.

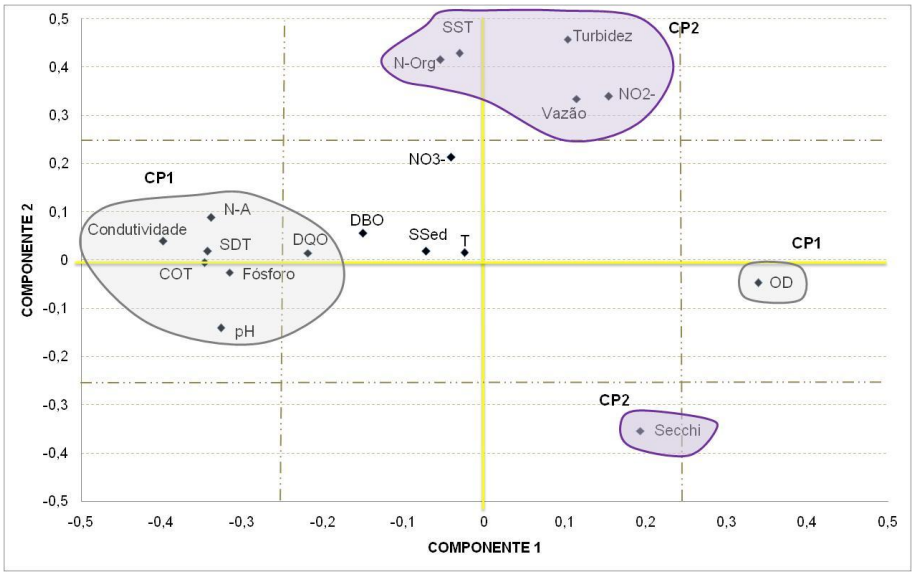


FIGURA 4.3 – Pesos das variáveis nas componentes principais 1 e 2

c) Escores das Componentes Principais

A Figura 4.4 apresenta os escores para as componentes 1 e 2 (os valores em forma de tabela encontram-se no Apêndice III). Os escores foram calculados substituindo-se as variáveis (parâmetros de qualidade de água) pelas coletas, também denominadas observações (linhas da Tabela 3.1), nas combinações lineares das componentes principais. Assim, estes 34 pontos representam o resultado do cálculo do valor das componentes 1 e 2 considerando-se as 34 coletas. Observa-se na Figura 4.4, a formação de alguns grupos. O grupo I é formado pelas amostras de 1 a 6, as quais se referem ao ponto de monitoramento P1 (ver Quadro 4.1). O ponto P1 é conhecido por apresentar amostras de água com melhor qualidade. Assim, uma possível interpretação é que as amostras pertencentes ao grupo IV, localizadas no extremo oposto do grupo I, sejam as que apresentam pior qualidade. O grupo

II é formado por coletas dos pontos P3, P4, P5 e P6 todas referentes às coletas ocorridas no dia 19/10/2005.

A amostra 7, referente ao ponto P2 e coletada no dia 10/08/2005, pode ser considerada um ponto *outlier*, estando afastada das demais. Provavelmente, um indicativo de que neste dia ocorreu algo incomum. Ao se analisar a Amostra I (Tabela 3.1), observou-se que realmente neste dia ocorreu a maior vazão para o ponto P2, sendo aproximadamente 5 vezes maior que as demais. O grupo III representa coletas nas quais as variáveis pertencentes às componentes 1 e 2 não exerceram muita influência.

O significado de se ter amostras próximas entre si, indicando agrupamentos, é que estas são semelhantes, podendo sugerir uma mesma característica de parâmetros de qualidade de água para as respectivas condições de amostragem, implicando do ponto de vista estatístico um mesmo “retrato” da condição de poluição.

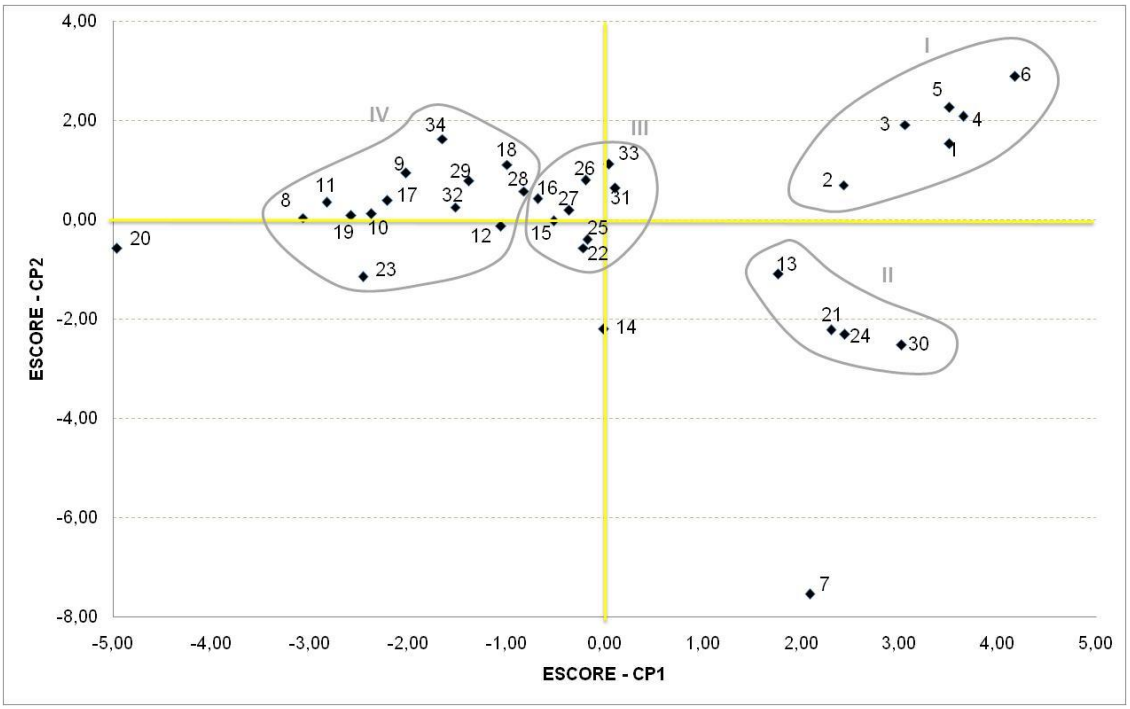


FIGURA 4.4 – Escores CP1 X CP2

4.1.4 Análise Fatorial

Neste item será realizada a análise fatorial da Amostra I. Os procedimentos para aplicação deste método consistem em: (1) verificação da normalidade, (2) aplicação do teste de Esfericidade de Bartlett e verificação da Medida de Adequacidade da Amostra de KMO e (3) de acordo com o resultado de (2), escolher o método a ser considerando para estimação dos fatores de acordo com o resultado de (1). No MATLAB, são realizadas as etapas (1) e (2), no STATISTICA, a etapa (3), que é a análise propriamente dita. Objetiva-se com a

aplicação deste método, comparar seus resultados, ou então combiná-los, com os resultados obtidos pela ACP.

a) Verificação da Normalidade Multivariada

A verificação da normalidade multivariada dos dados é necessária, visto que para utilização do método da máxima verossimilhança, pressupõe-se que os dados apresentem distribuição normal. Na verdade, a não verificação de normalidade não implica na total impossibilidade do uso do método de máxima verossimilhança, mas sim na confiabilidade dos resultados da análise.

O método utilizado foi proposto no item 2.1.2.2, seguindo-se os passos descritos para a avaliação da normalidade multivariada. A Figura 4.5 tem como referência as 34 coletas de água e relaciona o quadrado da distância generalizada e o qui-quadrado respectivo (os valores dos cálculos encontram-se no Apêndice IV). A verificação foi realizada no *software* MATLAB versão 5.3, através da função programada **normult**[13].

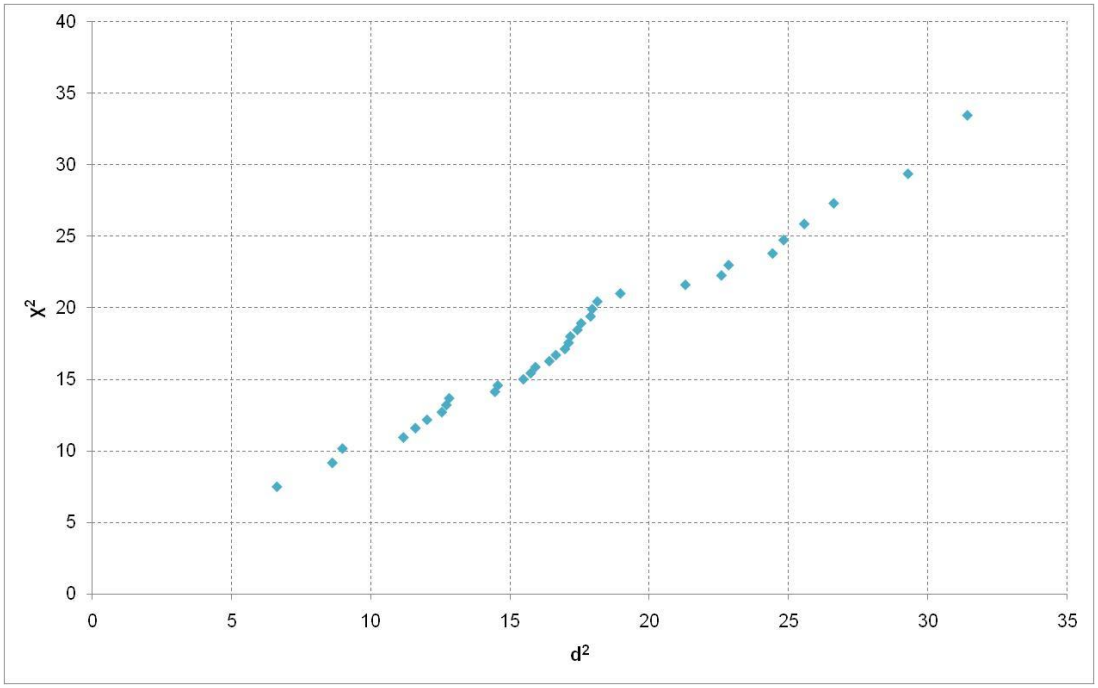


FIGURA 4.5 – Verificação da normalidade multivariada

Observa-se no gráfico que a distribuição dos pontos aproxima-se de uma reta e assim os dados observados podem ser considerados normais multivariados. Deste modo, será possível utilizar tanto o método de máxima verossimilhança para estimação dos fatores como o método das componentes principais, que não requer que os dados sejam normalmente distribuídos.

[13] Código fonte da função normult está descrito no Anexo II

b) Teste de Esfericidade de Bartlett e Verificação da Medida de Adequacidade da Amostra de KMO

Utilizou-se o teste de esfericidade de Bartlett (item 2.3.1) e a medida de adequacidade de Kaiser-Meyer-Olkin (item 2.3.2) para verificar se a análise fatorial era adequada à estrutura da matriz de dados. Os resultados obtidos foram:

Teste de Esfericidade de Bartlett:

χ^2 = 468,5864 com p-valor = 0 (com ν = 153), onde ν refere-se ao grau de liberdade

Medida de adequacidade da amostra de Kaiser-Meyer-Olkin (MSA)*:*

O resultado para o índice MSA foi 0,5476. Conforme exposto, valores altos (entre 0,5 e 1,0) indicam que a análise fatorial é apropriada. Com os resultados obtidos (p-valor=0 e MSA=0,5476), portanto, tem-se que a análise fatorial é adequada para as 18 variáveis.

c) Análise Fatorial: Método das Componentes Principais X Método da Máxima Verossimilhança

Neste item serão apresentados os resultados da análise fatorial realizada a partir dos métodos das componentes principais e da máxima verossimilhança. Para tanto se utilizou o software STATISTICA versão 6.0. A escolha deste *software* deve-se à possibilidade de realizar a rotação varimax, que é a rotação dos fatores (item 2.3.8) a qual tem por objetivo obter pesos altos para cada variável em um único fator e pesos baixos ou moderados nos demais fatores.

c.1) Estimação do Número de Fatores

A estimação do número de fatores foi determinada pelo critério de Kaiser (KAISER, 1958). Nesta análise, 5 fatores explicaram aproximadamente 78% da variância da amostra para o método das componentes principais e 71% para o método da máxima verossimilhança. A Tabela 4.6 apresenta os autovalores dos fatores e suas respectivas explicações da variância total e acumulada.

TABELA 4.6 – Autovalores e variância total

Fator	Autovalores		Variância Total Explicada(%)		Variância Total Acumulada (%)	
	CP	MV	CP	MV	CP	MV
1	5,40	4,75	30,02	26,38	30,02	26,38
2	3,46	2,85	19,25	15,84	49,27	42,22
3	2,26	1,85	12,53	10,25	61,81	52,47
4	1,70	1,95	9,47	10,83	71,28	63,30
5	1,26	1,35	7,00	7,48	78,27	70,78

c.2) Pesos dos 5 Fatores da matriz 34 X18

A Tabela 4.7 exibe os pesos (ou carregamentos) dos 5 fatores após ser realizada a rotação varimax. Os pesos com valor absoluto superior a 0,7 foram destacados.

TABELA 4.7 – Matriz dos pesos das variáveis nos fatores

	Fator 1		Fator 2		Fator 3		Fator 4		Fator 5	
Variável	CP	MV	CP	MV	CP	MV	CP	MV	CP	MV
DQO	0,34	0,34	0,10	0,18	**0,70**	0,04	0,23	-0,15	-0,14	0,60
DBO_5	0,27	0,25	-0,04	-0,14	**0,80**	-0,04	-0,19	0,08	0,12	**0,79**
SDT	**0,80**	**0,77**	0,06	0,05	0,03	0,05	0,09	0,04	-0,02	0
SST	0,01	0,05	**0,75**	-0,24	0,23	**0,72**	-0,25	-0,30	-0,37	0,20
SSed	0,10	0,12	-0,04	-0,02	0,13	-0,02	-0,02	**-0,83**	**-0,92**	0,17
N-A	**0,80**	**0,82**	0,22	0,07	0,10	0,22	0,07	0,49	0,44	0,10
N-Org	0,10	0,11	**0,86**	-0,06	-0,01	**0,85**	-0,09	0,19	0,17	-0,03
NO_2^-	-0,22	-0,24	0,18	**-0,88**	0,13	0,23	**-0,92**	0	0	0,17
NO_3^-	0,07	0,09	0,64	0,12	-0,27	0,43	0,23	0,25	0,31	-0,26
FÓSF	**0,70**	0,65	0,02	0,18	0,17	0,05	0,18	0,02	0,11	0,13
COT	**0,72**	0,67	0,05	0,19	0,31	0,05	0,20	-0,24	-0,29	0,25
TURB	-0,26	-0,27	**0,83**	-0,23	0,09	**0,88**	-0,29	0,04	0,03	0,08
COND	**0,93**	**0,93**	0,06	0,05	0,13	0,09	0,04	-0,04	-0,03	0,15
T	0,22	0,20	0,02	-0,10	**-0,72**	-0,03	-0,11	0,17	0,20	-0,53
OD	**-0,88**	**-0,88**	0,10	0,28	0,05	0,12	0,27	0,17	0,16	0,03
SECCHI	-0,45	-0,4	-0,63	0,19	0,05	-0,55	0,19	0,29	0,29	0,07
pH	**0,78**	**0,76**	-0,13	0,30	-0,30	-0,12	0,33	-0,08	-0,11	-0,26
Q	-0,06	-0,07	0,22	**-0,90**	0,30	0,21	**-0,87**	-0,07	-0,08	-0,31

A Tabela 4.8 mostra a composição de cada um dos 5 fatores . As variáveis que não obtiveram peso maior ou igual a l0,7l em nenhum dos fatores foram: NO_3^- e profundidade Secchi. Observa-se, ainda nesta tabela, que as variáveis fósforo, DQO, COT e T, presentes nos fatores resultantes da ACP, não se encontraram entre as variáveis presentes nos fatores obtidos pela MV. Além disso, a importância das variáveis também foi alterada de acordo com o método. O NO_2^- e a Q, por exemplo, pertencem ao F2 pelo método da máxima verossimilhança, que explica 15,84% da variância total, mas pelo método das componentes principais "caem" para o F4, que explica 9,47%, possuindo em termos gerais uma significância menor.

TABELA 4.8 – Composição dos 5 fatores

	COMPONENTE PRINCIPAL	MÁXIMA VEROSSIMILHANÇA
Fator 1	SDT(+), N-A (+), Fósforo(+),COT(+), Cond (+),pH (+), OD (-)	SDT(+), N-A (+), Cond (+), pH (+), OD (-)
Fator 2	SST (+), N-Org (+), Turb (+)	NO_2^- (-), Q (-)
Fator 3	DQO(+), DBO_5 (+), T (-)	SST (+), N-Org (+), Turb (+)
Fator 4	NO_2^- (-), Q (-)	SSed (-)
Fator 5	SSed (-)	DBO_5

c.3) Comunalidades

Em seguida, foram calculadas as comunalidades (item 2.3.4), que representam a porção de variância das variáveis distribuídas pelos fatores. Os valores são apresentados na Tabela 4.9. As comunalidades são importantes porque definem o critério de descarte de variáveis, ou seja, variáveis que apresentarem comunalidade inferior a 0,7 - valor adotado neste trabalho - poderão ser dispensadas.

TABELA 4.9 - Comunalidades

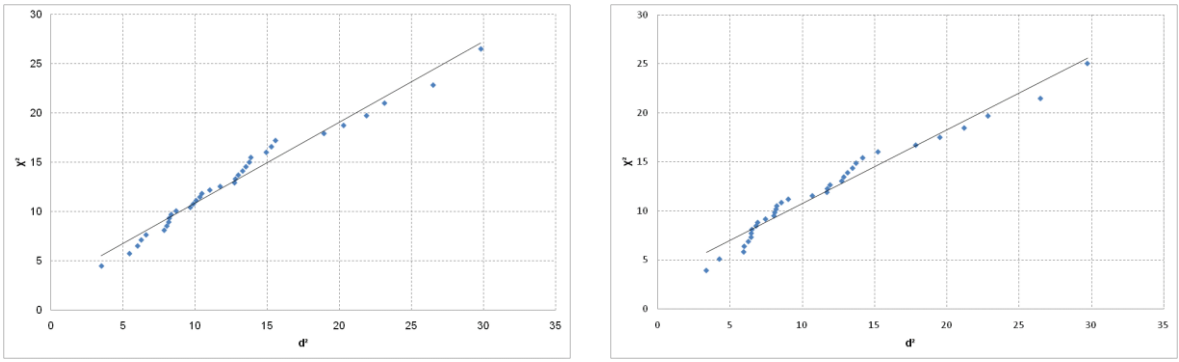
Variáveis	CP	MV
DQO	**0,69**	**0,53**
DBO_5	0,77	0,71
SDT	**0,65**	**0,60**
SST	0,81	0,71
SSed	0,88	0,73
N-A	0,90	0,97
N-Org	0,79	0,78
NO_2^-	0,94	0,92
NO_3^-	**0,63**	**0,34**
FÓSF	**0,57**	**0,48**
COT	0,74	**0,61**
TURB	0,85	0,91
COND	0,88	0,89
T	**0,62**	**0,36**
OD	0,89	0,90
SECCHI	0,72	**0,59**
pH	0,83	0,76
Q	0,91	0,95

Assim, de acordo com o critério estabelecido, 5 variáveis poderiam ser eliminadas pelo método das Componentes Principais: DQO, SDT, Nitrato, Fósforo e Temperatura da Água, e, 7 pelo método de máxima verossimilhança, excluindo além das anteriores a variável profundidade Secchi e o COT.

Neste caso, sugere-se que seja realizada uma nova estimação dos fatores considerando as 13 e 11 variáveis restantes para os métodos das componentes principais e da máxima verossimilhança, respectivamente.

c.4) Nova Verificação da Normalidade Multivariada

Realiza-se novamente a verificação da normalidade multivariada, visto que o método da máxima verossimilhança requer que a distribuição dos dados seja normal assim como se pressupõe para o Teste de Bartlett. A Figura 4.6 mostra que os pontos tendem a formar uma reta, comprovando a normalidade multivariada.



a) Método das CPs (13 variáveis) b) Método da MV (11 variáveis)

FIGURA 4.6 – Nova verificação da normalidade multivariada

c.5) Novo Teste de Esfericidade de Bartlett e Nova Verificação da Medida de Adequacidade da Amostra de KMO

Foram realizados novamente os testes de Bartlett e de KMO para as matrizes 34X13 e 34X11, referentes às variáveis que restaram (13 para o método das CPs e 11 para o da MV). Conforme mostra a Tabela 4.10, a análise fatorial para as 13 variáveis restantes, considerando-se o método das componentes principais, foi possível, visto que p-valor = 0 e MSA > 0,5. No entanto, para a matriz 34X11, resultante do método de máxima verossimilhança, embora p-valor = 0, MSA < 0,5, optando-se então por não prosseguir a análise por este método.

TABELA 4.10 – Novos testes de Bartlett e KMO

Método	Variáveis Analisadas	Teste de Esfericidade de Bartlett			Medida de Adequacidade da Amostra de Kaiser-Meyer-Olkin
		χ^2	p-valor	v (grau de liberdade)	MSA
CP	13	328,5001	0	78	0,525
MV	12	301,5117	0	66	0,4923

c.6) Nova Estimação do Número de Fatores

Novamente resultaram 5 fatores pelo critério de Kaiser. A variância total acumulada explicada pelos 5 fatores foi de aproximadamente 87%, ou seja, o novo modelo apresentou, neste aspecto, uma melhora em relação ao modelo anterior. A Tabela 4.11 apresenta os autovalores e suas respectivas explicações da variância total acumulada.

TABELA 4.11 – Autovalores e variância total

Fator	Autovalores	Variância Total Explicada(%)	Variância Total Acumulada (%)
1	4,03	31,02	31,02
2	3,35	25,79	56,81
3	1,53	11,77	68,58
4	1,26	9,70	78,29
5	1,14	8,80	87,08

c.7) Pesos dos 5 Fatores da matriz 34 X13

A Tabela 4.12 apresenta a matriz de pesos para os novos 5 fatores após realizada a rotação varimax. Como feito anteriormente, os pesos com valores absolutos superiores ou iguais a 0,7 foram destacados. Observou-se que o COT e a profundidade Secchi apresentaram pesos inferiores a 0,7 em todos os fatores, indicando não pertencer a nenhum destes primeiros 5 fatores, podendo estar inclusos em um dos outros 8 fatores restantes.

TABELA 4.12 – Matriz dos pesos das variáveis nos fatores

Variável	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
DBO_5	0,18	0,01	0,03	-0,01	**0,95**
SST	0,03	**0,78**	0,36	-0,20	0,16
SSED	0,12	0,02	**0,94**	0,01	0,04
N-A	**0,78**	0,21	-0,47	0,14	0,21
N-Org	0,12	**0,89**	-0,25	-0,03	-0,05
NO_2^-	-0,22	0,24	0,01	**-0,87**	0,25
COT	0,68	0,11	0,30	0,30	0,25
TURB	-0,27	**0,88**	-0,08	-0,23	0,02
COND	**0,92**	0,09	0,01	0,11	0,17
OD	**-0,89**	0,12	-0,16	0,27	-0,10
SECCHI	-0,47	-0,62	-0,26	0,18	0,04
pH	**0,81**	-0,16	0,04	0,30	-0,27
Q	-0,01	0,21	0,02	**-0,94**	-0,19

A Tabela 4.13 exibe a constituição de cada um dos fatores:

TABELA 4.13 – Novos fatores

	Fator	Variância (%)	Variáveis com peso ≥ l0,7l
Fatores com Alta Variância	1	31,02	N-A (+),Cond (+), pH (+), OD (-)
	2	25,79	SST (+), N-Org (+), Turb (+)
Fatores com Baixa Variância	3	11,77	SSed (+)
	4	9,7	NO_2^- (-), Q (-)
	5	8,8	DBO_5 (+)

Segundo a Tabela 4.13, observando-se o fator 1, ficou clara a oposição entre o nitrogênio amoniacal e o oxigênio dissolvido, visto que a amônia provoca consumo de oxigênio ao ser oxidada biologicamente. No entanto, apesar de se esperar que o pH variasse no sentido oposto do nitrogênio amoniacal, em virtude do ambiente se tornar mais ácido, não foi o que ocorreu.

O segundo fator apresentou a mesma formação da componente principal 2, sendo constituído pelos SST, turbidez e nitrogênio orgânico.

Um resultado interessante foi que em nenhum dos 2 primeiros fatores, que juntos explicaram quase 57% da variância, apareceram os parâmetros de determinação de matéria orgânica. Os fatores 1 e 2 acabaram por focar mais nos parâmetros que medem os sólidos e o nitrogênio presentes no corpo hídrico e a conseqüência direta trazida por eles que é a redução do oxigênio dissolvido.

O fator 3 foi representado exclusivamente pelos sólidos sedimentáveis e o fator 5 pela DBO_5. O fator 4 novamente destaca a importância do nitrogênio na avaliação da qualidade do corpo hídrico. O fato de a vazão ter variado no mesmo sentido que o nitrito, foi visto como um indicativo de que a vazão não auxiliou na diluição, refletindo os efeitos da poluição difusa, ou, potencial “re-suspensão” dos sólidos antes sedimentados.

Outra forma de se visualizar os pesos das variáveis nos fatores 1 e 2 é apresentada na Figura 4.7.

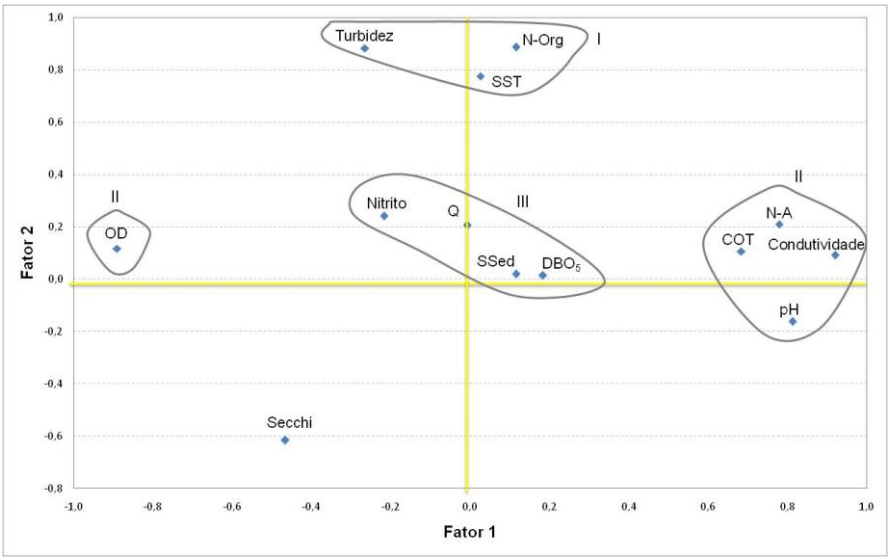


FIGURA 4.7 – Pesos das variáveis nos fatores 1 e 2

Observa-se a formação dos grupos I, II e III. O grupo III, próximo do zero do gráfico, indica que as variáveis nele contidas possuem pesos muito baixos tanto no fator 1 como no 2. O grupo I é formado pelas variáveis que constituem o fator 2 e o grupo II, com pesos altos na abscissa, formam o fator 1. A variável Secchi, no quadrante esquerdo inferior, apresenta pesos intermediários nos fatores 1 e 2, mas não chega a $\left|0{,}7\right|$.

c.8) Novas Comunalidades

Como pode ser observado na Tabela 4.14, todas as variáveis apresentaram comunalidade superior ou igual a 0,7, indicando que não há necessidade de eliminação de qualquer uma delas. A variável profundidade Secchi foi a que apresentou menor comunalidade sendo igual a 0,7.

TABELA 4.14 – Novas comunalidades

Variáveis	Comunalidades
DBO_5	0,94
SST	0,80
SSed	0,90
N-A	0,93
N-Org	0,87
NO_2^-	0,93
COT	0,72
TURB	0,91
COND	0,90
OD	0,91
SECCHI	0,70
pH	0,85
Q	0,96

c.9) Escores

A Figura 4.8 apresenta os escores para os fatores 1 e 2. Os valores em forma de tabela encontram-se no Apêndice V. Os dias e os pontos de monitoramento nos quais foram realizadas as 34 coletas foram apresentados no Quadro 4.1.

Nota-se que novamente a coleta 7, coletada no ponto P2 e no dia 10/08/2005, pode ser considerada um ponto *outlier*, estando afastada das demais. Neste dia, observou-se a maior vazão para o ponto P2, sendo aproximadamente 5 vezes maior que as demais, o que explica este “afastamento”.

Para as coletas 15, 18, 21, 24, 25, 26, 27, 28, 30, 31 e 33, que apresentaram escores inferiores a l0,5l nos fatores 1 e 2, uma possível interpretação é que quando foram realizadas estas coletas as variáveis pertencentes aos fatores 1 e 2 (N-A, condutividade, pH, OD, SST, N-Org e turbidez) não se apresentaram muito relevantes. Estas coletas referem-se às realizadas nos pontos: P3 (dias 03-04-06 e 24-05-06), P4 (19-10-05), P5 (todas as coletas exceto a realizada no dia 21-06-06) e P6 (19-10-05, 26-04-06 e 07-06-06). O interessante neste caso foi a presença de praticamente todas as coletas realizadas no P5 nesta faixa de escore, levantando a suspeita que talvez neste ponto de monitoramento as variáveis citadas acima não tenham tanto significado para avaliação da qualidade da água.

As coletas 1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 13, 16, 17, 19, 20, 22, 23, 29, 32, 34 apresentaram escores absolutos no fator 1 superiores aos seus escores no fator 2. O que indica que neste caso as variáveis N-A, condutividade, pH e OD apresentaram-se mais relevantes nestas coletas. Estas coletas referem-se aos pontos de monitoramento: P1 (todas as coletas), P2 (todas as coletas, exceto a 7), P3 (19-10-05 e 10-04, 26-04, 07-06, 21-06 de 2006), P4 (03-04 e 07-06 de 2006), P5 (21-06-06) e P6 (24-5 e 21-06 de 2006).

As coletas 12 e 14, que se referem às coletas dos dias 20-07-05 e 14-03-06, apresentaram escores absolutos no fator 2 maiores que no fator 1, indicando uma importância maior das variáveis SST, N-Org e turbidez nestas coletas.

No intuito de identificar quais seriam as amostras que representaram o melhor e o pior estado de qualidade de água da bacia, seguiu-se a seguinte interpretação:

- As amostras 1, 2, 3, 4, 5 e 6 referiram-se a coletas realizadas no P1, ponto de monitoramento que apresenta melhores condições de qualidade de água entre os demais. Assim, ratificando esta informação, estas amostras apresentaram escores negativos no fator 1 e no fator 2, o que indicou que para o fator 1, as concentrações de oxigênio se sobressaíram indicando melhor qualidade da água, e, para o fator 2, a concentração de sólidos foi mais baixa.

- A amostra 7, referente ao ponto P2 e coletada no dia 10/08/2005, pode ser considerada um ponto *outlier*, estando afastada das demais. Neste dia, observou-se a maior vazão para o ponto P2, sendo aproximadamente 5 vezes maior que as demais, o que explica este “afastamento”.

- A amostra 20 apresentou escores altos nos dois fatores, indicando que foi influenciada tanto pela poluição devida aos sólidos como pela poluição dos esgotos domésticos. Pode-se afirmar que esta amostra representou, então, a pior qualidade de água da bacia.

Deste modo, pode-se dizer que a qualidade da água das amostras possivelmente variou no sentido apresentado na Figura 4.8.

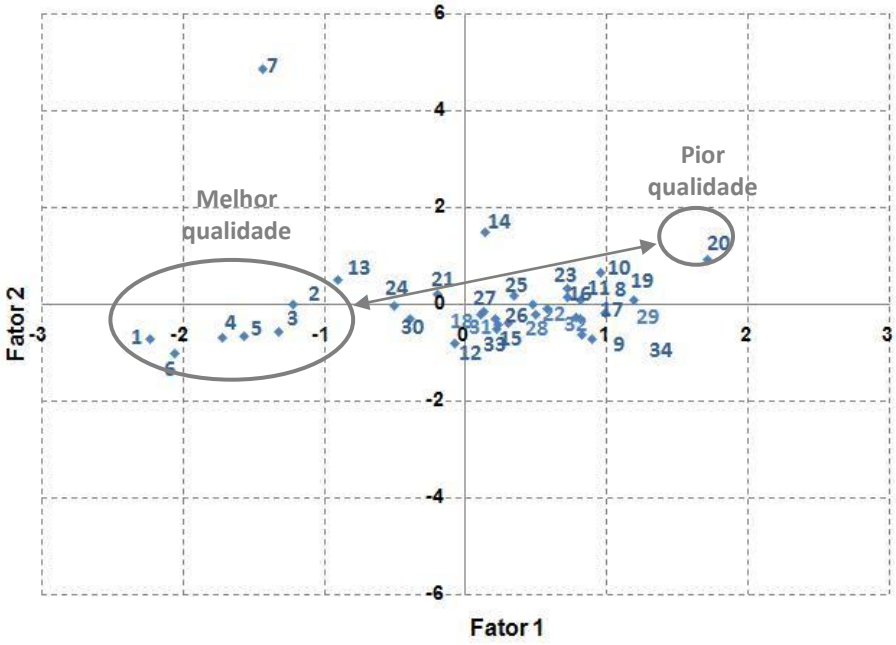


FIGURA 4.8 – Escores dos fatores 1 e 2

c.10) Matriz de Resíduos

A utilidade da matriz de resíduos é auxiliar na observação do ajuste do modelo. A equação 2.70, no item 2.3.6, mostra como se procede ao cálculo dos resíduos. No *software* STATISTICA, o valor *default a*dotado é igual a 0,10, considerando que a existência de resíduos maiores que 0,1 indicam que o modelo pode ser melhor ajustado. Visto que na Tabela 4.15 podem ser observados valores superiores a 0,10 (0,30 para a profundidade Secchi, por exemplo), pode-se afirmar que neste trabalho o modelo poderia ser melhor ajustado. Nota-se ainda que mesmo que se adotasse um critério mais brando que o “0,1” como “0,30”, por exemplo, o modelo ainda não estaria bem ajustado.

TABELA 4.15 – Matriz de resíduos

Variável	DBO_5	SST	SSed	N-A	N-Org	NO_2^-	COT	Turb	Cond	OD	Secchi	pH	Q
DBO_5	0,06	-0,05	0,01	-0,02	0,02	-0,02	-0,06	0,00	-0,02	0,02	-0,07	0,05	0,01
SST	-0,05	0,20	-0,04	0,05	-0,08	-0,03	-0,04	0,00	0,02	-0,05	0,12	-0,01	-0,01
SSed	0,01	-0,04	0,10	0,04	0,08	0,02	-0,09	0,01	0,04	0,00	0,08	0,02	-0,01
N-A	-0,02	0,05	0,04	0,07	-0,01	0,00	-0,03	-0,01	0,00	-0,01	0,06	-0,01	0,00
N-Org	0,02	-0,08	0,08	-0,01	0,13	0,00	-0,02	-0,05	0,02	0,00	0,03	-0,02	0,01
NO_2^-	-0,02	-0,03	0,02	0,00	0,00	0,07	0,03	0,03	0,03	0,04	0,04	0,04	-0,02
COT	-0,06	-0,04	-0,09	-0,03	-0,02	0,03	0,28	-0,01	-0,06	0,05	-0,02	-0,06	0,05
Turb	0,00	0,00	0,01	-0,01	-0,05	0,03	-0,01	0,09	0,03	0,00	0,05	0,05	-0,02
Cond	-0,02	0,02	0,04	0,00	0,02	0,03	-0,06	0,03	0,10	0,00	0,09	0,00	-0,03
OD	0,02	-0,05	0,00	-0,01	0,00	0,04	0,05	0,00	0,00	0,09	-0,03	0,05	0,02
Secchi	-0,07	0,12	0,08	0,06	0,03	0,04	-0,02	0,05	0,09	-0,03	0,30	0,02	-0,01
pH	0,05	-0,01	0,02	-0,01	-0,02	0,04	-0,06	0,05	0,00	0,05	0,02	0,15	0,00
Q	0,01	-0,01	-0,01	0,00	0,01	-0,02	0,05	-0,02	-0,03	0,02	-0,01	0,00	0,04

4.1.5 Análise de Agrupamentos

A análise de agrupamentos da Amostra I (Tabela 3.1) foi realizada através do método de agrupamento hierárquico (item 2.4.2) no *software* STATISTICA, considerando como variáveis as coletas (linhas da matriz), de modo a agrupar as coletas de amostras de água do rio, visando encontrar quais refletiam melhor e pior qualidade do corpo hídrico.

Primeiramente foram calculadas as correlações cofenéticas (item 2.4.3) através da função programada **cophenet**[14] no *software* MATLAB, visando identificar qual seria o melhor tipo de ligação e de distância a serem adotados nesta análise, considerando para tanto a correlação mais próxima de 1. Assim, a melhor correlação obtida foi igual a 0,6598, referente à distância euclidiana e ligação média (Tabela 4.16). Para a distância *Mahalanobis*, as correlações calculadas resultaram em números imaginários, sendo descartadas.

TABELA 4.16 – Correlação cofenética para a Amostra I - Coletas

	LIGAÇÃO				
DISTÂNCIAS	*Simples*	*Completa*	*Média*	*Centróide*	*Ward*
Euclidiana	0,6082	0,6568	0,6598	0,6082	0,5148
Quadrado da Dist. Euclidiana	0,5852	0,6160	0,6186	0,6173	0,4807
Cityblock	0,6077	0,5577	0,5971	0,5957	0,5032
Mahalanobis	-	-	-	-	-

[14] O algoritmo desta função é apresentado no Anexo IV

Os agrupamentos das variáveis resultantes constam no seguinte dendrograma (Figura 4.9):

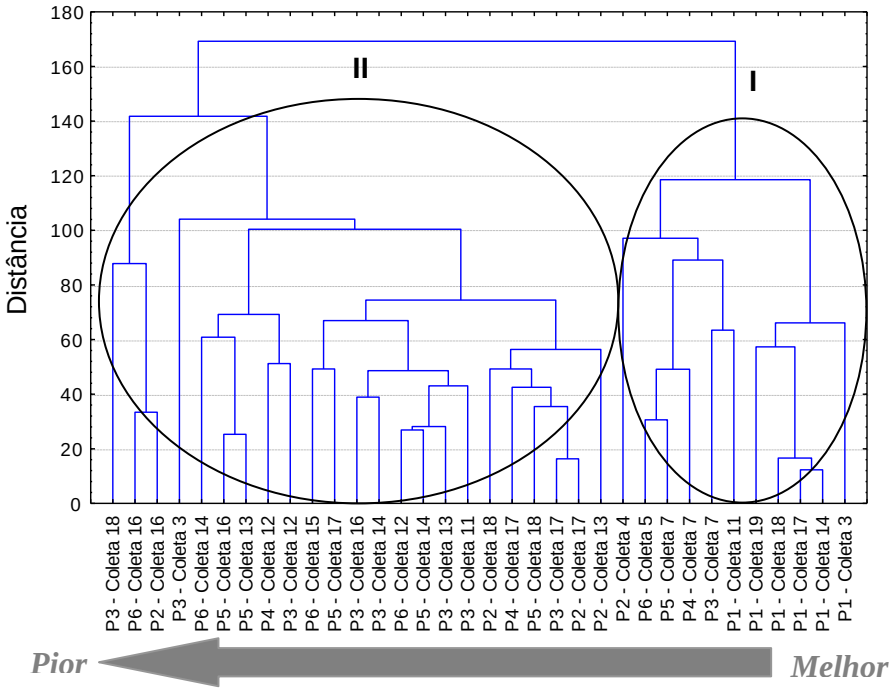


FIGURA 4.9 – Dendrograma da Amostra I - Coletas

Nota-se na Figura 4.9 a formação de dois agrupamentos principais. No **agrupamento I** ficaram agrupadas todas as coletas referentes ao ponto de monitoramento P1, conhecido por estar localizado em uma área menos poluída da bacia; as coletas 7 dos pontos P2, P4, P5 e P6; e as coletas 4 do ponto P2 e 5 do ponto P6. Estas amostras pertencentes ao **agrupamento I**, embora em pontos de monitoramento mais poluídos, foram coletadas em períodos em que o rio apresentava vazão maior, auxiliando na diluição dos poluentes (ver Tabela 3.1).

No **agrupamento II**, ficaram reunidas as coletas que apresentaram qualidade da água intermediária ou pior, ficando assim as coletas divididas em dois grupos gerais: coletas de água com boa qualidade (**I**) e coletas de água com média e baixa qualidade (**II**).

Deste modo, pode-se afirmar que a qualidade da água piora no sentido apresentado na Figura 4.9, com as amostras que refletem melhor qualidade no lado direito do dendrograma e as que refletem pior qualidade no lado oposto.

A Tabela 4.17 apresenta o histórico de agrupamento das 34 variáveis (coletas).

TABELA 4.17- Histórico do agrupamento das 34 variáveis

Passo	Distância	Nº de Grupos	Grupos Unidos
1	12,31	33	{P1 - Coleta 14} e {P1 - Coleta 17}
2	16,32	32	{P2 - Coleta 17} e {P3 - Coleta 17}
3	16,60	31	Grupos unidos no passo 1 e {P1 - Coleta 18}
4	25,32	30	{P5 - Coleta 13} e {P5 - Coleta 16}
5	26,87	29	{P5 - Coleta 14} e {P6 - Coleta 12}
6	28,14	28	Grupos unidos no passo 5 e {P3 - Coleta 13}
7	30,59	27	{P5 - Coleta 7} e {P6 - Coleta5}
8	33,39	26	{P2 - Coleta 16} e {P6 - Coleta 16}
9	35,46	25	Grupos unidos no passo 2 e {P5 - Coleta 18}
10	38,92	24	{P3 - Coleta 14} e {P3 - Coleta 16}
11	42,53	23	Grupos unidos no passo 9 e {P4 - Coleta 17}
12	43,04	22	Grupos unidos no passo 6 e {P3 - Coleta 11}
13	48,61	21	Grupos unidos no passos 12 e 10
14	49,15	20	Grupos unidos no passo 7 e {P4 - Coleta 7}
15	49,23	19	Grupos unidos no passo 11 e {P2 - Coleta 18}
16	49,26	18	{P5 - Coleta 17} e {P6 - Coleta 15}
17	51,19	17	{P3 - Coleta 12} e {P4 - Coleta 12}
18	56,37	16	Grupos unidos no passo 15 e {P2 - Coleta 13}
19	57,31	15	Grupos unidos no passo 3 e {P1 - Coleta 19}
20	60,85	14	Grupos unidos no passo 4 e {P6 - Coleta 14}
21	63,46	13	{P1 - Coleta 11} e {P3 - Coleta 7}
22	66,11	12	Grupos unidos no passo 19 e {P1 - Coleta 3}
23	66,94	11	Grupos unidos nos passos 13 e 16
24	69,23	10	Grupos unidos nos passos 17 e 20
25	74,47	9	Grupos unidos nos passos 18 e 23
26	87,84	8	Grupos unidos no passo 8 e {P3 - Coleta 18}
27	89,13	7	Grupos unidos nos passos 14 e 21
28	97,09	6	Grupos unidos no passo 27 e {P2 - Coleta 4}
29	100,38	5	Grupos unidos nos passos 24 e 25
30	104,09	4	Grupos unidos no passo 29 e {P3 - Coleta 3}
31	118,55	3	Grupos unidos nos passos 22 e 28
32	141,78	2	Grupos unidos nos passos 29 e 30
33	169,23	1	Grupos unidos nos passos 31 e 32

4.2 ANÁLISE DOS PONTOS DE MONITORAMENTO DA BACIA DO ALTO IGUAÇU

Nesta análise, as variáveis avaliadas foram os pontos de monitoramento P1, P2, P3, P4, P5 e P6. O objetivo foi identificar quais destes seis pontos poderiam ser considerados os mais representativos na avaliação da qualidade de água do Alto Iguaçu. Para tanto, trabalhou-se com as medianas dos valores de cada um dos parâmetros de qualidade de água em cada uma das estações, resultando na Tabela 3.2**,** uma matriz 6 X 18, denominada Amostra II.

4.2.1 Estatística Descritiva das 6 variáveis

A Tabela 4.18 apresenta a média, o desvio padrão, a variância e o coeficiente variação de cada uma das 6 variáveis originais: P1, P2, P3, P4, P5 e P6. As variáveis foram dispostas em ordem decrescente do valor do coeficiente de variação.

TABELA 4.18 – Estatística descritiva das 6 variáveis

Variáveis	***Média***	***Desvio Padrão***	***Variância***	***Coeficiente de Variação***
P2	28,47	47,99	2303,04	1,69
P6	22,86	38,45	1478,19	1,68
P5	26,48	42,56	1811,33	1,61
P3	25,29	39,73	1578,34	1,57
P4	23,16	34,28	1175,45	1,48
P1	10,36	15,34	235,23	1,48

Nota-se que os pontos de monitoramento P2 e P6 apresentaram o maior coeficiente de variação e, portanto, maior grau de dispersão dos dados, enquanto que os pontos P4 e P1 apresentaram o menor valor para o coeficiente de variação: 1,48. No entanto, mesmo sendo o menor valor do coeficiente de variação, este não pode ser considerado baixo, ou seja, este valor não indica que os pontos P4 e P1 apresentaram um baixo grau de dispersão.

No geral, o alto grau de dispersão representado pelos coeficientes de variação já era aguardado, visto que neste caso as medidas estatísticas foram realizadas considerando-se a mediana de diversos parâmetros de qualidade de água, medidos em diferentes unidades, os quais apresentaram valores em torno de zero (por exemplo, sólidos sedimentáveis), mas também valores muito altos, em torno de 200 (por exemplo, sólidos dissolvidos totais) conforme apresentado na Tabela 3.2.

4.2.2 Matriz de Correlação para as 6 variáveis

A Tabela 4.19 exibe as correlações existentes entre as 6 variáveis referentes à Amostra II. Os valores em vermelho são aqueles superiores ou iguais a 0,5. Nota-se que o ponto de monitoramento P1 apresentou correlações muito baixas com os outros pontos de monitoramento, com exceção do ponto P5. Em contrapartida, os pontos de monitoramento P2, P3, P4, P5 e P6 são altamente correlacionados entre si. Os resultados acabaram evidenciando a diferenciação dos pontos de monitoramento no que tange à qualidade de água, assim, o ponto P1 acabou se distanciando dos demais por apresentar melhores condições de qualidade de água, por estar justamente inserido em uma área de manancial.

TABELA 4.19 – Matriz de correlação para as 6 variáveis

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1	1,00	0,36	0,45	0,34	0,5	0,45
P2	0,36	1,00	0,99	0,87	0,97	0,97
P3	0,45	0,99	1,00	0,89	0,99	0,99
P4	0,34	0,87	0,89	1,00	0,93	0,94
P5	0,5	0,97	0,99	0,93	1,00	0,99
P6	0,45	0,97	0,99	0,94	0,99	1,00

4.2.3 Análise de Componentes Principais dos Pontos de Monitoramento

Para a avaliação dos pontos de monitoramento, realizou-se a análise de componentes principais, com o intuito de se conhecer as relações existentes entre eles e a relevância de cada um deles na avaliação da qualidade da água da bacia do Alto Iguaçu.

a) Estimação do Número de Componentes Principais

Para a estimação do número de componentes principais, primeiramente obteve-se a matriz de correlação da Amostra II (Tabela 3.2). Calcularam-se, em seguida, os autovalores e autovetores correspondentes desta matriz. A importância das componentes principais é definida por seus autovalores, assim quanto maior for o autovalor, mais relevante será a componente principal. Os autovalores correspondem à variância explicada por cada uma das componentes principais. A Tabela 4.20 apresenta os autovalores e a variância total explicada pelas componentes principais.

TABELA 4.20 – Autovalores e Variância Total Explicada

Componente Principal	*Autovalor*	*Variância Explicada (%)*	*Variância Explicada Acumulada (%)*
1	5,04	83,93	83,93
2	0,80	13,29	97,22
3	0,15	2,53	99,74
4	0,01	0,17	99,91
5	0,00	0,06	99,97
6	0,00	0,03	100,00

Para a seleção do número de componentes principais, adotando-se o critério de Kaiser (KAISER, 1958), seria retida apenas a primeira componente, visto que somente esta apresentou autovalor maior que 1. Contudo, considerando o método "*Scree-Plot"* de Cattell (1966), apresentado no item 2.2.4, seria possível trabalhar ainda com a segunda componente, visto que a soma da variância explicada das duas primeiras componentes resultou em aproximadamente 97%, que é um valor bastante satisfatório (Figura 4.10). Além disso, mais informação será agregada quando da interpretação dos resultados, enquanto que a contribuição das últimas quatro componentes principais seria marginal.

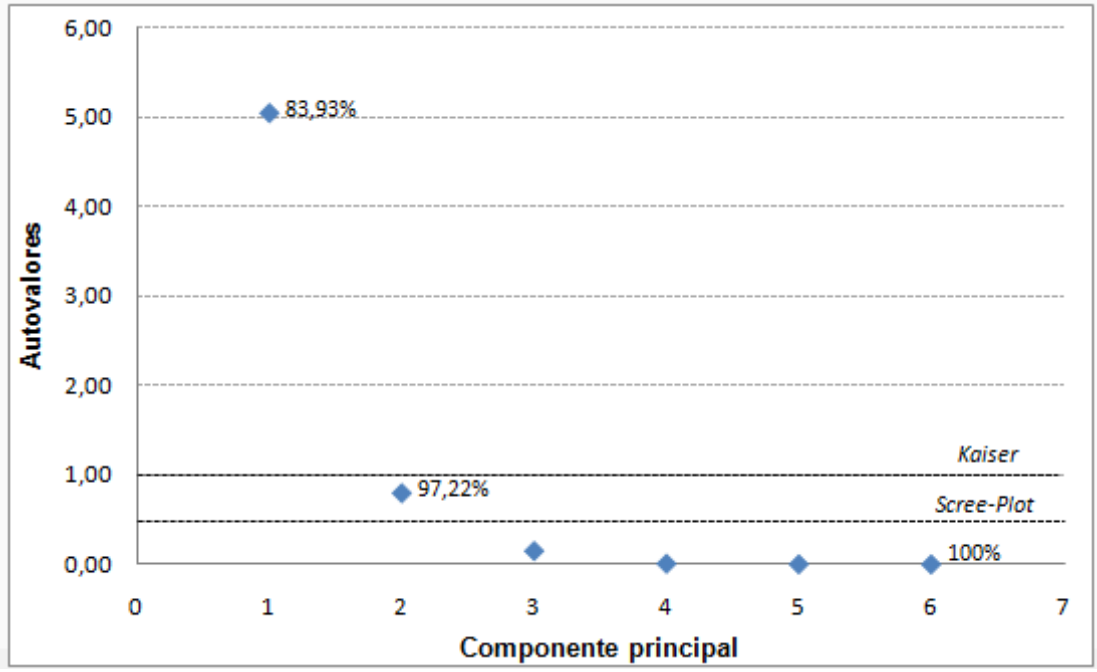


FIGURA 4.10 – Seleção do número de componentes principais

b) Componentes principais da Amostra II

A Figura 4.11 mostra a importância de cada umas da variáveis originais (pontos de monitoramento) nas componentes principais 1 e 2. A importância das variáveis originais nas componentes principais é marcada pelos pesos que as variáveis originais têm na combinação linear que define a componente principal. Os pesos ou *loadings* são na verdade os autovetores, que são ordenados de acordo com os seus respectivos autovalores, estes em ordem decrescente. Assim como os pesos, o cálculo do coeficiente de correlação entre variáveis e componentes principais (ver equação 2.36) também é importante para averiguar a relevância das variáveis originais nas componentes principais, auxiliando na interpretação dos resultados. Na Figura 4.11, são apresentados os pesos (em azul) e as correlações (em vermelho) de cada uma das variáveis originais nas componentes principais 1 e 2. As variáveis com correlações maiores ou iguais a 0,7 – em valores absolutos – são consideradas relevantes para a definição das componentes principais.

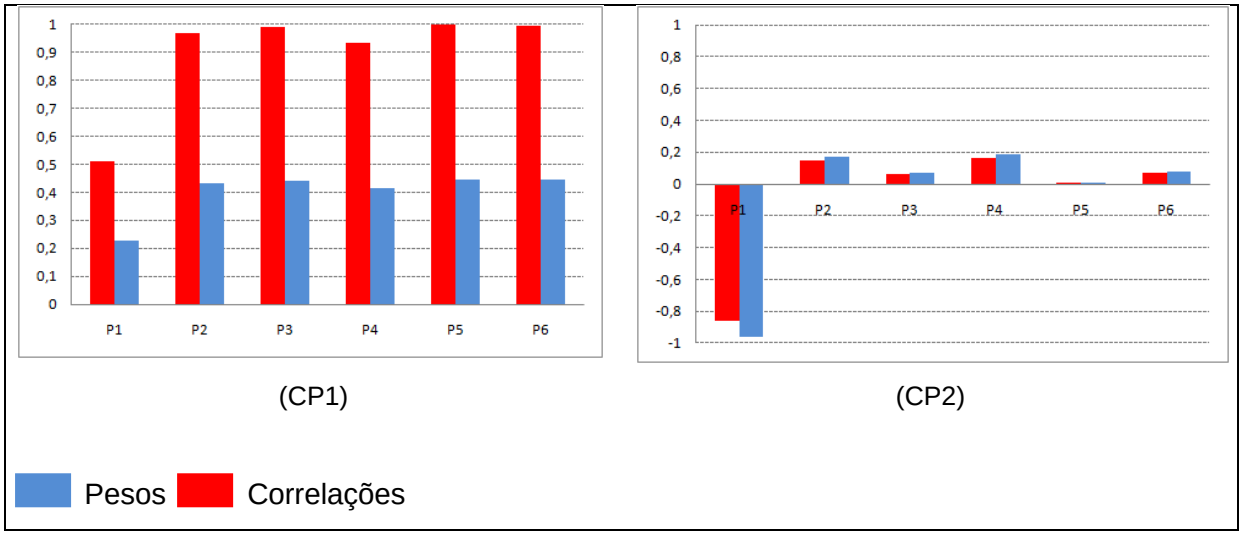


FIGURA 4.11 – Pesos e correlações das variáveis originais

Observa-se na figura anterior que as variáveis originais que mais se destacaram na primeira componente principal foram os pontos de monitoramento P2, P3, P4, P5 e P6, enquanto que na segunda componente principal a variável com maior destaque foi o ponto de monitoramento P1. Na primeira componente principal, pode-se dizer que os pontos P2, P3, P4, P5 e P6 apresentaram praticamente o mesmo peso para a definição da CP1. Já a CP2 mostra a oposição entre o ponto de monitoramento P1 e os demais, o que faz sentido, visto que o ponto P1 está localizado em uma área de manancial e apresenta amostras de água com melhor qualidade, além disso, o ponto P1 foi a variável com maior peso nesta componente.

Uma forma de visualizar os resultados obtidos é a representação gráfica apresentada na Figura 4.12. Nota-se que os pontos de monitoramento P2, P3, P4, P5 e P6 praticamente se sobrepõem, com pesos mais representativos na CP1 e peso quase nulo na CP2. O ponto P1, em contrapartida, apresenta peso menos significativo na CP1, e, bastante significativo na CP2, especialmente quando comparado aos outros pontos de monitoramento.

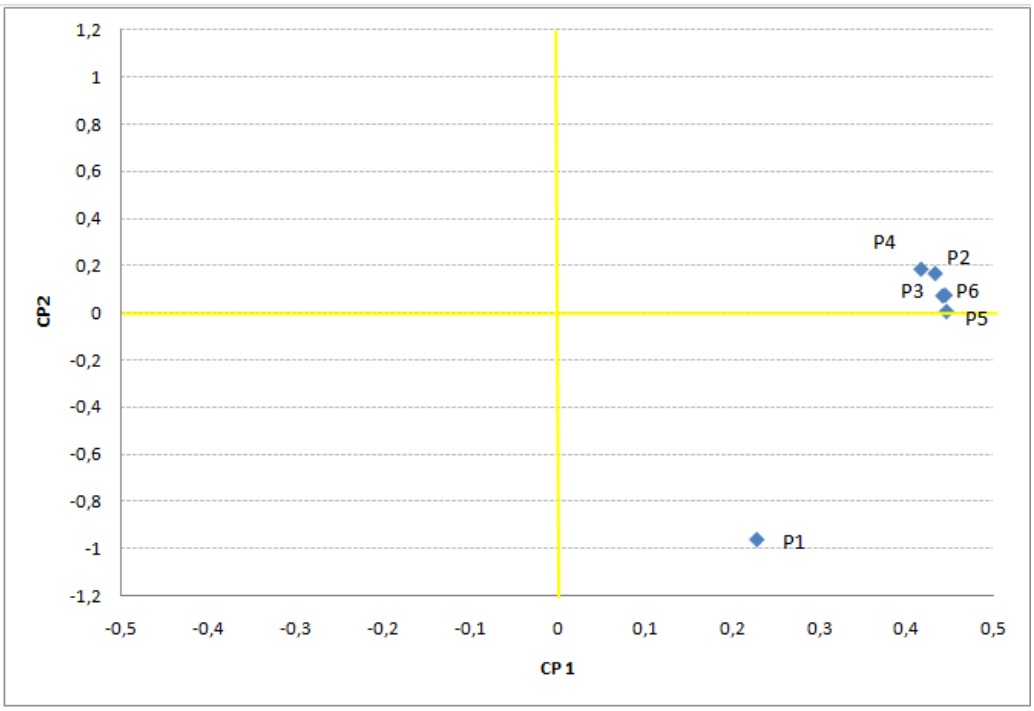


FIGURA 4.12 – Peso das variáveis nas componentes principais

Quanto à relevância dos pontos de monitoramento, em virtude de a primeira componente principal explicar cerca de 84% da variância da amostra, pode-se afirmar que os pontos de monitoramento P2, P3, P4, P5 e P6 foram os que mais se destacaram, por possuírem maior peso nesta componente. Ou seja, o resultado da análise acabou tendendo para os pontos de monitoramento que freqüentemente apresentam amostras de água com qualidade indesejável. Contudo, deve-se atentar que apesar de a participação do ponto P1 na primeira componente principal ser inferior às demais, esta não deve ser considerada desprezível.

TABELA 4.21 – Pesos das variáveis originais na CP1 e na CP2

	CP1	CP2
P1	0,2279	-0,9621
P2	0,432	0,1684
P3	0,4406	0,0734
P4	0,4158	0,1865
P5	0,4449	0,0068
P6	0,4433	0,0758

4.3 SÍNTESE DOS RESULTADOS

Neste capítulo foram apresentados os resultados e as discussões para as duas análises realizadas: Análise Global da Bacia do Alto Iguaçu e Análise dos Pontos de Monitoramento. Na primeira análise, considerou-se a Amostra I (Tabela 3.1), na qual as variáveis foram os parâmetros de qualidade de água e as observações foram as amostras de água. Na segunda, referente à Amostra II (Tabela 3.2), as variáveis foram os pontos de monitoramento da bacia.

Na Análise Global, os resultados da Análise de Componentes Principais mostraram que as variáveis que mais se destacaram, podendo ser interpretadas como as mais relevantes na avaliação da qualidade da água da bacia, foram: OD, SDT, Nitrogênio Amoniacal, Fósforo, COT, Condutividade, pH, SST, Nitrogênio Orgânico e Turbidez. Estas variáveis foram consideradas de maior importância por estarem nas componentes principais 1 e 2, que juntas explicaram aproximadamente 50% da variância total da amostra.

A Análise Fatorial da Amostra I foi realizada utilizando-se tanto o método das componentes principais como da máxima verossimilhança. Os valores das comunalidades indicaram que algumas variáveis poderiam ser excluídas da análise: DQO, SDT, NO_3^-, Fósforo e Temperatura - pelo método das CPs; e complementarmente, todas estas, mais a profundidade Secchi e o COT, com base no método da máxima verossimilhança. No entanto, quando realizados os testes de Bartlett e o cálculo da medida de adequacidade da amostra (MSA) de KMO, o valor de MSA foi inferior a “0,5”. Assim, optou-se por prosseguir a análise somente pelo método das componentes principais. A composição dos novos dois primeiros fatores, que juntos explicaram 56,81% da variância total, ficou do seguinte modo: OD, Nitrogênio Amoniacal, Condutividade, pH, SST, Nitrogênio Orgânico e Turbidez.

Assim, comparando-se os resultados obtidos pela ACP e pela AF, observa-se que pela AF descartaram-se ainda três variáveis consideradas como mais relevantes pela ACP:

SDT, Fósforo e COT. O COT, em especial, não foi eliminado em razão da comunalidade, mas por não apresentar peso superior a $\left|0,7\right|$ em nenhum dos 5 primeiros fatores.

Portanto, as variáveis consideradas como mais relevantes foram: OD, Nitrogênio Amoniacal, Condutividade, pH, SST, Nitrogênio Orgânico e Turbidez, as quais ilustram os aspectos de degradação da matéria orgânica e sua interação com a dinâmica de transporte de sólidos. Um resultado interessante, é que os parâmetros de determinação de matéria orgânica – DBO_5, DQO e COT – tidos normalmente como de grande importância na avaliação qualitativa dos corpos hídricos não apareceram nos fatores 1 e 2, que possuem alta variância. Caso se adotasse um critério um pouco mais brando, no entanto, o COT apareceria no 1º fator (Tabela 4.12), refletindo possivelmente que, em termos gerais, este teste sofre menos interferências que o da DBO, apresentando resultados mais robustos.

A Análise de Agrupamentos da Amostra I foi realizada primeiramente para os parâmetros de qualidade de água e, posteriormente, para as coletas amostradas. No primeiro caso, a análise resultou em 3 agrupamentos: **Agrupamento I** [Q, pH, N-A, OD, N-Org, Fósforo, NO_2^-, NO_3^-, SSed, Temperatura, Turbidez, COT e DBO_5]; **Agrupamento II** [Secchi, SST e DQO] e **Agrupamento III** [Condutividade e SDT]. Para a análise das coletas, estas se dividiram basicamente em 2 grupos: no **Agrupamento I**, reuniram-se as coletas que refletiram melhor estado da qualidade do corpo hídrico, incluindo todas as amostragens realizadas no P1 - localizado em uma área menos degradada da bacia – e outras amostradas em outros pontos de monitoramento, mas em dias que apresentaram maior vazão. O **Agrupamento II** foi formado pelas demais coletas de amostras d'água, consideradas de qualidade inferior.

Para a Análise dos Pontos de Monitoramento da Bacia do Alto Iguaçu (item 4.2), referente à Amostra II, na qual as variáveis passaram a ser os pontos de monitoramento e as observações as medianas dos parâmetros de qualidade de água, realizou-se apenas a Análise de Componentes Principais. Na primeira componente principal, os pontos P2, P3, P4, P5 e P6 foram os que apresentaram correlações e pesos mais representativos para a sua definição. Já a CP2 mostrou a oposição entre o ponto de monitoramento P1 e os demais, o que pode ser explicado em razão de o ponto P1 estar localizado em uma área de manancial e apresentar amostras de água com melhor qualidade, além disso, o ponto P1 foi a variável com maior peso nesta componente. Quanto à relevância dos pontos de monitoramento, o resultado da análise acabou tendendo para os pontos de monitoramento que freqüentemente apresentam amostras de água com qualidade indesejável (P2, P3, P4, P5 e P6). Contudo, deve-se atentar que apesar de a participação do ponto P1 na primeira componente principal ser inferior às demais, esta não foi considerada desprezível.

CAPÍTULO V

5. CONCLUSÕES E RECOMENDAÇÕES

5.1 CONCLUSÕES

Atualmente, graças à tecnologia dos computadores pessoais e ao grande número de *softwares* comerciais disponíveis, a teoria da análise multivariada transformou-se em uma ferramenta mais acessível, ganhando campo em diversas áreas como: Psicologia, Ciências Sociais e Biológicas, Educação, Ergonomia, Física, Química, Geologia, Engenharia, etc. (MINGOTI, 2005). E, não foi diferente para área ambiental, no que tange à gestão de qualidade da água.

Na gestão de recursos hídricos e, por conseguinte, na gestão da qualidade da água de um determinado rio, muitos esforços têm sido despendidos para compreensão da complexa interação entre aspectos qualitativos e quantitativos. A base para tal fundamentação esta na necessária consistência entre as séries históricas hidrológicas (quantitativas) e os dados de monitoramento de qualidade da água.

As limitações das séries de dados qualitativos da água já são consensuais considerando: (i) as dificuldades em se realizar campanhas sistemáticas; (ii) as séries históricas de qualidade de água, quando existentes, encontram-se defasadas ou não dependentes das séries hidrológicas; (iii) a dificuldade em se definir quais de fato são os parâmetros de qualidade de água mais relevantes para uma dada região / bacia hidrográfica / rio.

Esta pesquisa visou esclarecer estes fatos dado que a aplicação encaixada da técnica não foi verificada em muitos artigos da literatura. Por exemplo, para utilização apropriada de algumas técnicas estatísticas multivariadas, faz-se necessária a realização de testes prévios como a verificação da distribuição normal multivariada, o teste de esfericidade de Bartlett e a medida de adequacidade da amostra de Kaiser-Meyer-Olkin. Além disso, deve-se atentar também sobre o número de variáveis (p) e observações (n), visto que quando “n > p”, o número de dados disponíveis (graus de liberdade) é maior, provendo soluções mais estáveis.

No caso da análise de componentes principais, esta pode ser vista como uma maneira objetiva de se obter índices. Mas, na área de qualidade de água, interpretar as componentes principais desta maneira não seria trivial, visto que os parâmetros de qualidade de água não variam necessariamente no mesmo sentido, por exemplo, o OD

reflete melhores condições de qualidade quando em altas concentrações, a DBO e a DQO em altas concentrações demonstram exatamente o contrário.

No presente estudo, o ganho com a utilização da ACP não foi a obtenção de índices, mas sim identificar as variáveis mais relevantes na avaliação da qualidade da água do Alto Iguaçu bem como a relação existente entre elas. Esta também foi a contribuição da AF, que simplifica ainda mais a estrutura vinda da ACP, diminuindo a contribuição das variáveis com menor significância e aumentando a contribuição das que possuem maior significância.

A Análise Global da Bacia do Alto Iguaçu avaliou 18 variáveis de qualidade de água, em uma primeira instância. Após a utilização do critério de redução de variáveis - as comunalidades - o modelo final passou a ter 13 variáveis para análise, considerando-se o método das componentes principais para estimação dos fatores. Destas 13 variáveis, 7 foram explicadas pelos 2 primeiros fatores, os quais apresentaram variância de aproximadamente 57%. Os outros 3 fatores, além de possuírem variâncias mais baixas, não agregaram informações vantajosas para este estudo.

Assim, as 7 variáveis de qualidade de água, que explicaram 57% da variância total, sendo então consideradas as mais relevantes para a bacia do Alto Iguaçu, foram: OD, Nitrogênio Amoniacal e Nitrogênio Orgânico, Condutividade, pH, SST e Turbidez, as quais ilustram os aspectos de degradação da matéria orgânica e sua interação com a dinâmica de transporte de sólidos. A presença das variáveis nitrogênio amoniacal e nitrogênio orgânico entre as mais relevantes revela ainda que as amostras de água foram coletadas em pontos onde o foco de poluição se encontrava próximo. Note-se que não foram considerados relevantes os parâmetros de determinação de matéria orgânica – DBO_5, DQO e COT, o que indicou que os resultados das análises estatísticas focaram no impacto, por exemplo nas alterações dos valores de pH e de oxigênio dissolvido, e não nos de efeito, como as respostas da DBO em razão da variação do oxigênio dissolvido.

Quanto ao agrupamento das coletas das amostras do rio, obtiveram-se dois grupos principais: os de coletas que refletiram melhor qualidade do corpo hídrico, formado principalmente por coletas realizadas no ponto P1, próximo a uma área de manancial da bacia; e o outro formado por grande parte das outras coletas, as quais refletiram o estado de degradação do rio, evidenciando e confirmando que em sua totalidade, a qualidade da água da bacia apresenta-se inadequada.

Na Análise dos Pontos de Monitoramento da Bacia do Alto Iguaçu, as variáveis tornaram-se os próprios pontos de monitoramento, representados pelas medianas dos parâmetros de qualidade de água, visando encontrar quais estações de monitoramento seriam mais representativas para avaliação da qualidade do corpo hídrico bem como a relação existente entre elas. Os pontos que se mostraram mais relevantes foram P2, P3,

P4, P5 e P6, os quais definiram a primeira componente principal, que isoladamente já explicava cerca de 84% da variância total. Contudo, com a agregação da segunda componente principal a variância explicada subiu para cerca de 97%. Considerar a segunda componente principal nos resultados foi importante para identificar a oposição entre os pontos de monitoramento. Na CP2, o ponto de monitoramento que mais se destacou foi o P1 que não foi abrangido pela CP1, além disso, os valores do peso e da correlação de P1 com a CP2 foram contrários aos dos outros pontos de monitoramento, o que faz sentido, visto que o ponto P1 está localizado em uma área de manancial e apresenta amostras de água com melhor qualidade.

Quanto aos resultados obtidos, deve-se ressaltar, no entanto, a necessidade de comparação com resultados de análises futuras, para que se tenha segurança quanto à confiabilidade destas afirmações.

A principal contribuição deste trabalho, mais ainda do que os resultados obtidos em si, está relacionada à sistemática aplicação das técnicas multivariadas, contrariando aplicações similares descritas na literatura; complementarmente introduzir a visão de planejamento e consenso, requeridas para a adequada implementação dos instrumentos de gestão de recursos hídricos, como condição de contorno à aplicação dos métodos. Descobriu-se que o emprego deste tipo de análise pode não ser tão complexo quanto dar significância aos seus resultados, lembrando sempre, da importância da execução de testes prévios. Assim, se fosse necessário dar uma nota para o patamar em qual se encontra hoje a utilização da análise multivariada para a gestão da qualidade de água no Brasil e a escala fosse um iceberg, poderia dizer-se que se está apenas em sua parte visível. Deste modo, há a ainda a necessidade de se “descobrir” a parte “não visível” do potencial das técnicas dentro das inúmeras possibilidades de uso da análise multivariadas e estratégias de avaliação,

O intuito deste trabalho foi incentivar o uso das técnicas multivariadas - considerando a relevância de seus resultados e reconhecendo estas técnicas como poderosas ferramentas estatísticas - visando elucidar as interações existentes entre as variáveis que estruturam a gestão de qualidade de água, ainda desconhecidas por nós. Os resultados aqui obtidos compilam a primeira experiência do uso da análise multivariada para a gestão da qualidade de água no Rio Iguaçu na Região Metropolitana de Curitiba.

5.2 RECOMENDAÇÕES

Os conhecimentos produzidos a partir deste trabalho representam o ponto de partida para a aplicação da análise estatística multivariada na avaliação dos dados de monitoramento de qualidade de água e de vazão da bacia do Alto Iguaçu. Faz-se necessário

ainda realizar algumas complementações que visam o aprofundamento desta dissertação bem como a confirmação de seus resultados. Algumas recomendações referem-se a:

- Desenvolver técnicas para análise de consistência de dados de qualidade de água similares às técnicas de preenchimentos de falhas para as séries hidrológicas, visando avaliar o impacto de não se desprezar dados obtidos de procedimentos amostrais tradicionais e, que em geral, são caros e complexos;

- Dar continuidade às campanhas de monitoramento de qualidade de água da bacia do Alto Iguaçu, visto que para aplicação da análise multivariada, à princípio, é interessante que o número de observações seja maior que o número de variáveis avaliadas;

- Realizar novas análises considerando uma base de dados maior, de forma a complementar os resultados aqui obtidos;

- Desempenhar uma avaliação mais profunda dos resultados obtidos para o agrupamento das amostras (item 4.1.5), investigando-se quais eram as condições hidro-climatológicas durante os procedimentos de coleta de amostras, procurando pela existência de padrões;

- Incluir a variável “precipitação (mm)” em estudos futuros, buscando sua relação com a vazão e os demais parâmetros de qualidade de água, além de outras variáveis não contempladas;

- Realizar a ACP e a AF para cada ponto de monitoramento individualmente, visando encontrar quais são os parâmetros de QA mais importantes para cada um dos pontos monitorados;

- Separar os dados de acordo com os períodos de cheia e estiagem e realizar novas análises para estes 2 conjuntos de dados, procurando investigar se os parâmetros de QA mais relevantes na bacia diferem em razão das características hidrológicas.

REFERÊNCIAS

APHA. **Standard Methods for the Examination of Water and Wastewater**. 20 ed. USA, Washington : APHA, 1998.

BENGRAÏNE, K. & MARHABA, T.F. **Using principal component analysis to monitor spatial and temporal changes in water quality**. Journal of Hazardous Materials, p. 179-195, 2003.

BRITO, L.T.L.; SRINIVASAN, V.S.; SILVA, A.S.; GALVÃO, C.O.; RIBEIRO, P.H.B. **Variabilidade da qualidade da água do rio Salitre.** Anais do 4º Simpósio Brasileiro de Captação e Manejo de Água de Chuva, Bahia, Julho, 2003.

CATTELL, R.B. **The screen test for the number of factors**. Multivariate Behavioral Research, 1, p. 140-161, 1966.

CATTELL, R.B. & JASPERS, J. **A general plasmode (No. 30-10-5-2) for factor analytic exercises and research**. Mult. Behav. Res. Monogr. 67, p.1 – 212, 1967.

CETESB. **Variáveis de qualidade das águas**. Companhia de Tecnologia de Saneamento Ambiental. [http://www.cetesb.sp.gov.br/Agua/rios/variaveis.asp, Acesso em:01/03/2009].

CHIGUTI, M. **Aplicação da análise multivariada na caracterização dos municípios paranaenses segundo suas produções agrícolas**. Curitiba-PR. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, 2005.

CONAMA. **Resolução nº 357 de 17 de março de 2005**. Dispõe sobre a classificação e diretrizes ambientais para o enquadramento dos corpos de água superficiais, bem como estabelece as condições e padrões de lançamento de efluentes. Relator: Marina Silva. Diário Oficial da União, Brasília, 18 de março de 2005.

DIXON, W. & CHISWELL, B. **Review of aquatic monitoring program design**. Water Resources , nº 30, p. 1935-1948, 1996.

GROSSMAN, G.D., NICKERSON, D.M. & FREEMAN, D.M. **Principal component analysis of assemblage structure data: utility of tests based on eigenvalues**. Ecology, 72, p. 341-347, 1991.

HAIR JR, J.F.; ANDERSON, R.E.; TATHAM, R.L. **Multivariate data analysis**. New York: Editora Maxwell MacMillan International Editions, 1987.

HAIR JR., J. F. *et al.* **Análise Multivariada de Dados**. 5 ed. Tradução: Adonai Schlup Sant'nna e Anselmo Chaves Neto. Porto Alegre: Bookman, 2005. Tradução de: Multivariate Analysis.

HARDYCK, C.D. & PETRINOVICH, L.F. **Introduction to Statistics for the Behavioral Science**s. 2ª ed. Philadelphia: Saunders. 1976.

JOHNSON, D.E. **Applied multivariate methods for data analysis**. Brooks/Cole Publishing Company, 1998.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 4 ed. New Jersey: Prentice Hall, 1998.

KAISER, H.F. **The varimax criterion for analytic rotation in factor analysis**. Psychometrika, 23, p. 187- 200, 1958.

KAISER, H.F. and RICE, J. **Little Jiffy Mark IV**. Educational and Psychological Measurment, 34 (Spring), p. 111-117,1974.

KNAPIK, H. G. **Modelagem da Qualidade da Água na Bacia do Alto Iguaçu:Monitoramento e Calibração**. Monografia de conclusão de curso (Engenharia Ambiental). Universidade Federal do Paraná, Curitiba, 130 f, 2006.

KNAPIK, H. G. et al. **Análise crítica da calibração do modelo de qualidade de água Qual2e – Estudo de caso da bacia do Alto Iguaçu**. Revista de Gestão da Água - REGA, Volume 5, nº 2, julho/dezembro/2008.

MARDIA, K. V.; KENT, J. T.; BIBBY J. M. **Multivariate Analysis**. London: Academic Press, Inc., 1979.

MARQUES, J. M. **Apostila de análise multivariada aplicada à pesquisa**. Universidade Federal do Paraná, Curitiba-PR, 2003.

MARQUES, M.A.M. **Aplicação da Análise multivariada no estudo da infra-estrutura dos serviços de saúde dos municípios paranaenses**. Curitiba-PR. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, 2006.

McGARIGAL, K.; CUSHMAN, S.; STAFFORD,S. **Multivariate statistics for wildlife and ecology research**. New York: Springer Verlag, 2000.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Editora UFMG, Belo Horizonte, 2005.

NONATO, E.A.; VIOLA, Z.G.G.; ALMEIDA, K.C.B.; SCHOR, H.H.R. **Tratamento estatístico dos parâmetros da qualidade das águas da bacia do Alto Curso do Rio das Velhas**. Química Nova, Vol. 30, Nº 4, p. 797-804, 2007.

OUYANG, Y. **Evaluation of river water quality monitoring stations by principal component analysis** . Water Research, 39, p. 2621-2635, 2005.

PORTO, M. F. A. **Prospecção de Pesquisa em Qualidade da Água**. Centro de Gestão de Estudos Estratégicos, Brasília, Brasil, 2003.

PORTO, M. F. A et al. **Bacias Críticas: Bases Técnicas para definição de metas progressivas para o seu enquadramento e integração com os demais sistemas de gestão - Estudo de caso da Bacia do Alto Iguaçu**. Curitiba: Universidade Federal do Paraná – Departamento de Hidráulica e Saneamento (FINEP/CT-HIDRO). Projeto concluído, 2007.

RENCHER, A.C. **Methods of multivariate analysis**. New York: John Wiley, 2002.

SHRESTHA, S.; KAZAMA, F. **Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan**. Environmental Modelling & Software, Vol. 22, p. 464-475, 2007.

SHARMA, S. **Applied Multivariate Techniques**. Ed. John Wiley and Sons. EUA, 1996.

SUDERHSA. 2000. **Plano de Despoluição Hídrica da Bacia do Alto Iguaçu**. Programa de Saneamento Ambiental da Região Metropolitana de Curitiba – Relatórios Finais. Curitiba: SUDERHSA. Projeto concluído.

TUCCI, C.E.M. **Gestão da água no Brasil**. Brasília: UNESCO, 2001.

VEGA, M.; PARDO, R.; BARRADO, E.; DEBÁN, L. **Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis**. Water Research, Vol. 32, Nº 12, p. 3581-3592, 1998.

VON SPERLING, M. 2005. **Introdução à Qualidade das Águas e ao Tratamento de Esgotos**. 3. ed. Minas Gerais: DESA / UFMG. 452 p.

WILLET, P. **Similarity and Clustering in Chemical Information Systems**. Research Studies Press, Wiley, New York, 1987.

WUNDERLIN, D.A.; DÍAZ, M.P; AMÉ, M.V.; PESCE, S.F.;HUED, A.C.; BISTONI, M..A.(2001). **Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality, a case study: Suquía River Basin (Córdoba – Argentina)**.Water Research, Vol. 35, Nº 12, p. 2881-2894, 2001.

YEUNG, I.M.H. **Multivariate analysis of the Hong Kong Victoria Harbour water quality data**. Environmental Monitoring and Assessment, Nº 59, p. 331-342, 1999.

YU, C.C.; QUINN, J.T.; DUFOURNAUD, C.M.; HARRINGTON, J.J.; ROGERS, P.P.; LOHANI, B.N. **Effective dimensionality of environmental indicators: a principal component analysis with bootstrap confidence intervals**. Journal of Environmental Management. 53: p. 101-119, 1998.

APÊNDICES

Apêndice I - Fotos dos pontos monitorados

Apêndice II – Dados de qualidade de água da Bacia do Alto Iguaçu

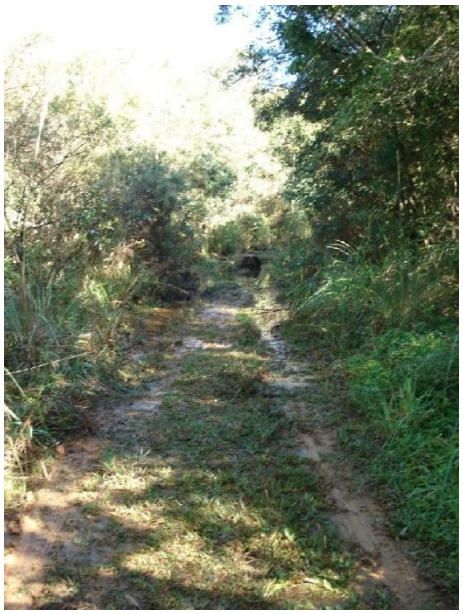
Apêndice III – Escores das componentes principais para Análise I

Apêndice IV – Quadrado das distâncias generalizadas e qui-quadrados respectivos – Análise I

Apêndice V - Escores dos novos 5 primeiros fatores – Análise I

APÊNDICE I - FOTOS DOS PONTOS MONITORADOS

OLARIA



(A)

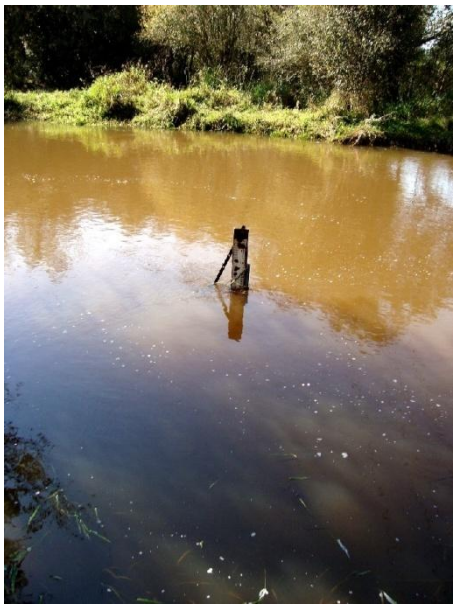
(B)

(C)

(D)

FIGURA A1. (A) Entrada da Olaria; (B) Equipe de campo; (C) Régua de nível e (D) Ponto de coleta

PR-415



(A)

(B)

FIGURA A2. (A) Régua de nível e (B) Ponte PR-415

P1

(A)

(B)

FIGURA A3. (A) Canal de água limpa e (B) Ponto de coleta

P2

(A)

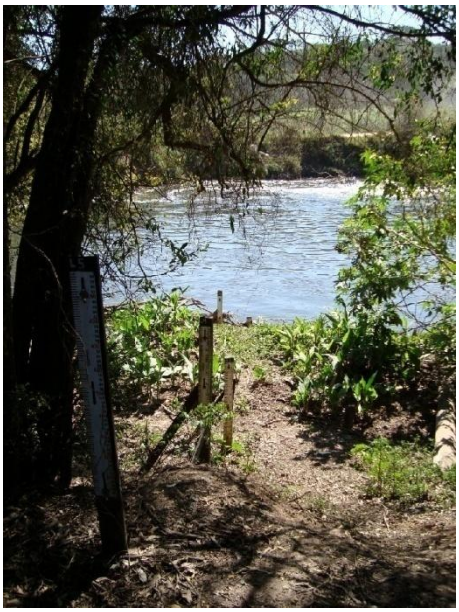
(B)

(C)

FIGURA A4. (A) BR-277; (B) Leitura da condutividade e pH; (C) Réguas de nível

P3

(A)



(B)

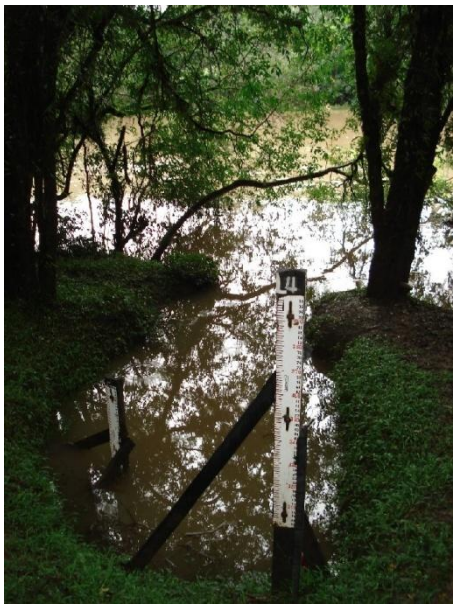
(C)

(D)

FIGURA A5. (A) Margem esquerda do Rio Iguaçu no P3; (B) Réguas de nível na margem esquerda; (C) Margem direita e (D) Coleta de amostra na margem direita

P4

(A)

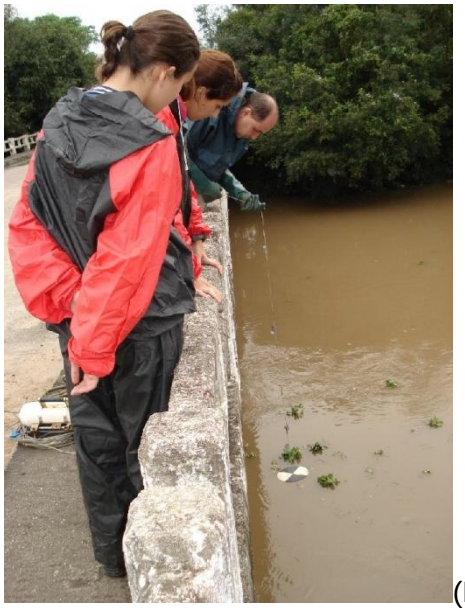


(B)

FIGURA A6. (A) Coleta no P4 e (B) Régua de nível

P5

(A)



(B)

FIGURA A7. (A) Réguas de nível e (B) Ponto de amostragem

P6

(A)

(B)

FIGURA A8. (A) Ponto de coleta e (B) Equipamentos de campo

APÊNDICE II – DADOS DE QUALIDADE DE ÁGUA DA BACIA DO ALTO IGUAÇU

As tabelas a seguir apresentam os dados de qualidade de água obtidos *in situ* e em laboratório para a bacia do Alto Iguaçu. As células em cinza referem-se à falta de dados que podem ter ocorrido em virtude da impossibilidade da medição, por falta de calibração de equipamentos ou defeito nos mesmos, quebra do pHmetro e imprevistos em laboratório, entre outros. As células em amarelo representam dados duvidosos, como por exemplo, em alguns casos onde a DBO é maior que a DQO, em outros onde a condutividade é muito inferior em relação aos outros dados da série e o nitrogênio amoniacal apresenta valores superiores a 40 mg/L. As células em verde referem-se a ensaios onde não se foi possível detectar valores, possivelmente em virtude do método utilizado e sua faixa de detecção.

TABELA A1 - ESTAÇÃO DE MONITORAMENTO OLARIA

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SSed (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	30/06/05													24,7	16,2	4,95		5,87	1,5
2	20/07/05													27,5	14,1	6,1		6,2	3,25
3	05/10/05																		6,24
4	19/10/05													30,1	18,8		90	6,175	4,04
5	31/10/05												17,5	27	17,4		90	6,09	8
6	14/11/05													38,2	19,9		80	6,7	3,15
7	15/12/05													27,3	20,3	5	60	6,67	2,52
8	14/03/06													29,7	22,8	4,28	55	6,77	2,52
9	03/04/06													44,5	19,4	5,58	60	6,607	0,84
10	10/04/06													26,9	20,6	5,74	60	6,81	1,59
11	26/04/06													37,8	19,5	6,94	75	6,965	3,86
12	11/05/06													31,1	16,4	4,09	65	6,58	1,74
13	24/05/06													40,3	13,9	8,72	10	6,8	4,04
14	07/06/06													27	16,7	6,84	80	6,9	4,7
15	21/06/06													27,3	16,2	6,8	60	6,82	5,44
16	19/07/06													43,1	14,3	4,52	70	6,84	3,4
17	13/03/08	23,00	2,71										19,45	8,1	21,1	3,8	70	5,93	2,4
18	16/04/08	16,00	1,97	66	10	< 0,1	1,20	0,76	0,070		0,177	9,06	12	10,14	17,8	3,35		5,52	1,44
19	07/05/08	14,00	< 2,00	56	2	< 0,1	0,22	1,01	0,019	0,319	0,051	4,55	12,55	5,8	13,9	4,74	100	6,67	8
20	04/06/08	23	4,23	114	40	< 0,3	0,22	1,46	0,022	0,552	0,127	6,88	96,5	7	14,8	3,18	30	5,857	3,98
21	20/08/08	20,32	3,96	74	22	< 0,1	0,45	2,58	0,014		0,105	7,01	24,8	53	18,45	4,99	110	7	3

TABELA A2 - ESTAÇÃO DE MONITORAMENTO P1

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SSed (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	17/06/05		49,2	40	9	<0,1	2,93	2,52	0,245	0,295	0,021	6,795	11,6	26,2	17,1	5	50	5,7	4,60
2	30/06/05	32	48	74	25	<0,1	2,52	2,52	0,070	0,263	0,011	5,528	7,53	27	16,55	6,05		6,1	3,73
3	20/07/05	64	24,0	14	22	<0,1	0,11	0,055	0,123	0,269	0,001	6,24	10,75	4,4	13,5	6,6	75	6,09	6,00
4	10/08/05	21,95	34,5	111	23	1,00	0,53	ND	0,040	0,357	0,024	7,79	28,67	22	12,0	7,5	30	6,13	12,32
5	14/09/05	15,09	9,04	71	29	<0,1	0,27	0,22	0,095	0,253	0,046	7,86	14,28	19,5	12,6			6,15	
6	05/10/05	13,59	23,28	123	8	<0,1	0,67	0,56	0,153	0,156	0,017	28,66	16,9	23,1	18,1			5,70	11,41
7	19/10/05	15,5	6,24	29	15	<0,1	0,05	0,52	0,113	0,344	ND	5,54	7,36	19,3	18,4	5,62	90	5,98	7,82
8	31/10/05	20,38	2,64	69	13	< 0,1			0,098	0,277	ND	7,014		17,8	17,6	6,4	100	5,75	16,31
9	14/11/05	16,66	7,2	31	15	<0,1	0,23	0,94	0,150	0,261	0,006	5,608	7,19	20,6	19,5		80	6,47	3,18
10	15/12/05	15,39	6,76	140	5	< 0,1	0,09	0,57	0,298	0,008	0,005	7,03	9,73	23,5	19,9			6,7	
11	14/03/06	12,31	6,12	97	20	<0,1	0,65	0,16	0,040	0,346	0,022	7,8	22,31	25,1	22,1	4,88	35	6,67	2,46
12	03/04/06	19,12	5,02	124	30	0,1	0,11	1,02	0,041	0,403	ND	10,03	23,46	29,3	20,1	4,74	45	6,63	1,71
13	10/04/06	7,97	4,6	161	25	<01	0,39	0,33	0,061	0,677	ND	6,94	13,51	26,4	20,2	5,74	60	6,67	1,93
14	26/04/06	11,15	4,44	22	12	<0,1	0,11	0,65	0,037	0,452	0,024	6,67	9,13	22,7	19,7	7,06	60	7,009	3,02
15	11/05/06	14,34	6	105	23	<0,1	0,77	0,05	0,113			5,94	11,64	11,5	16,2	3,12	65	6,88	2,55
16	24/05/06	11,29	9,94		25	<0,1	ND	0,05	0,027	0,069	0,040	5,53	11,9	19,5	14,7	6,64	70	6,9	2,52
17	07/06/06	14,6	1,98	21	14	< 0,1	0,11	0,28	0,044	0,384	0,018	5,45	9,94	18,7	16,6	7,44	70	6,7	2,4
18	21/06/06	13,00	3,32	9	7	< 0,1	0,16	0,16	0,029	0,096	0,019	8,24	12,32	19	15,7	7,2	60	6,72	2,48
19	19/07/06	6,66	2,58	60	8	< 0,1	0,33	0,38	0,031	0,047	0,035	3,62	12,1	16,6	15,06	7,52	100	6,56	2,48
20	13/03/08	23,00	2,75	78	< 0,0001	0,2	0,56	0,62	0,047	0,094	0,130	14,65	20,5	6,5	20,9	8,76		5,725	4,11
21	16/04/08	27,20	5,06	56	12	< 0,1	< 0,1	0,38	0,035		0,079	10,08	17,15	5,9	17,9	4,41	85	5,618	3,98
22	07/05/08	19,00	3,01	68	6	< 0,1	0,06	0,9	0,009	0,114	0,044	5,62	18,6	5	15	4,88	100	6,45	13,71
23	04/06/08	24	6,89	34	4	<0,2	<0,22	1,12		0,505	0,148	6,85	112	6	14,6	3,32	35	5,928	8,01
24	20/08/08	17,27	3,09	60	14	< 0,1	0,67	2,58	0,008		0,115	5,16	21,45	44	19	4,15	50	6,924	6,36

TABELA A3 - ESTAÇÃO DE MONITORAMENTO P2

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SSed (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	17/06/05		72	62	14	<0,1	41,83	22,68	0,871	0,274	0,223	7,749	9,13	126,5	17,6	3,2	50	6,4	11,00
2	30/06/05	49,6	51,6	163	21	<0,1	59,98	10,08	1,319	0,207	0,367	11,73	10,13	128,3	17,2	2,65		6,16	13,00
3	20/07/05	27,04	43,2	86	27	<0,1	6,82	1,16	0,380	0,188	0,354	11,12	16,16	111,4	12,5	3,3		6,73	11,00
4	10/08/05	28,22	23,0	86	87	<0,1	6,32	6,32	0,445	0,975	0,200	6,83	74,00	65,1	12,5	7,7	10	6,51	40,50
5	14/09/05	10,57	5,1	134	36	0,6	0,38	0,99	0,631	0,949	0,152	7,87	71,5	67,68	12,5		15	6,48	33,75
6	05/10/05	40,75	24,00	152	21	1,2	0,28	1,51	0,997	0,675	0,142	10,62	159	59,1	17,5		15	6,50	106,80
7	19/10/05	7,75	12,08	101	15	0,3	4,04	0,69	1,077	0,167	0,234	6,79	11,73	85,8	18,9	3,72	65	6,79	12,00
8	31/10/05	18,82	12	100	30	0,1			0,613	0,489	0,052	7,571	60	51,2	17,6	4,44	20	6,75	33,25
9	14/11/05	26,66	7,92	139	14	<0,1	4,90	1,6	0,666	0,202	0,218	25,46	8,06	106,7	21,6		80	7,2	11,25
10	15/12/05	30,78	18,88	207	26	0,1	6,73	0,85	0,324	0,357	0,896	9,26	11,73	137,9	20,4		50	7,26	10,04
11	14/03/06	21,54	16,32	162	15	<0,1	20,85	3,23	0,098	0,764	0,873	7,33	8,47	150,5	22,8	1,62	40	7,342	
12	03/04/06	15,94	9,1	183	12	0,1	7,81	1,07	0,071	0,319	0,399	10,51	12,36	141,3	21,3	1,92	50	7,21	
13	10/04/06	58,96	40,2	177	34	0,2	11,43	1,47	0,057	0,354	0,123	23,25	10,75	183,6	21,3	1,4	45	7,252	8,6
14	26/04/06	20,72	24,72	116	17	0,1	9,02	1,85	0,053	0,654	0,855	12,66	11,51	136,8	20,2	2,48	50	7,287	8,6
15	11/05/06	20,72	5	230	26	0,3	10,65	0,16	0,084		0,941	14,73	9,9	140,7	16,4	1,38	15	7,37	9,08
16	24/05/06	37,09	13,13	249	20	<0,1	8,62	0,55	0,029	0,870	0,518	13,58	4,4	141,7	15,4	1,86	45	7,3	8,6
17	07/06/06	38,52	17,7	149	32	< 0,1	8,85	1,34	0,048	0,186	1,821	18,15	12,83	163,1	17,2	2,32	30	7	7,76
18	21/06/06	64,19	24,9	143	18	0,4	1,70	1,26	0,083	0,076	1,501	24,59	14,87	171,1	16,5	2,02	15	7,28	7,55
19	19/07/06	34,11	14,55	245	19	0,3	15,26	3,95	0,021		1,535	45,9	16,53	182,12	15,15	0,9	15	7,1	6,92
20	13/03/08	26,00	8,59	177	124	< 1,5	1,90	2,24	0,170	0,379	0,627	12,86	87	14,5	21,1	10,3	20	6,719	37,5/ 43,5
21	16/04/08	32,00	6,26	196	32	< 0,4	5,35	2,04	0,296		0,503	9,15	45,5	32,6	18,5	2,62	30	6,113	11
22	07/05/08	17,00	5,31	102	32	< 0,4	3,7	1,57	0,101	0,569	0,256	7,19	38,05	16	15,6	4,63		6,81	14
23	04/06/08	31	4,74	124	42	< 0,4	1,34	2,35	0,116	0,916	0,236	8,17	162,5	20	15,2	3,11	30	6,646	34,5
24	20/08/08	21,34	7,58	128	34	1	8,85	3,25	0,099		0,365	6,03	50,5	205	18,75	2,63		7,246	12,25

TABELA A4 - ESTAÇÃO DE MONITORAMENTO P3

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SS (mL/L)	N-A (mg/L)	N-Org (mg/L)	N-K (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	17/06/05		81,6	7	31	<0,1	25,7	25,2	40,82	0,131	0,156	0,211	11,69	14,49	97,7	18,2	2,1	35	6,1	19,96
2	30/06/05	31,8	61,2	190	12	<0,1	61,49	7,56	69,05	0,662	0,150	0,483	14,55	9,56	138,9	17,85	0,95		6,26	13,00
3	20/07/05	41,6	92,4	181	27	<0,1	7,43	0,22	7,65	0,439	0,199	0,534	12,22	12,07	126,1	13,3	1,4	50	6,66	12,00
4	10/08/05	17,25	22,8	138	49	<0,1	7,37	6,58	13,95	0,472	1,024	0,114	7,00	56,00	76,7	13,1	7,8		6,67	74,70
5	14/09/05	16,6	15,16	110	29	0,1	1,04	0,05	1,09	0,659	0,777	0,135	6,5	41,14	63	13		20	6,44	87,23
6	05/10/05	16,6	20,16	159	11	0,1	1,12	0,84	1,96	1,061	0,664	0,250	6,69	65	57,4	17,9			6,46	78,34
7	19/10/05	24,8	9,6	81	48	0,4	1,15	1,5	2,65	0,337	0,191	0,200	9,74	20,6	72,1	19,1	3,48	45	6,59	24,59
8	31/10/05	15,68	8,64	91	33	< 0,1				0,699	0,495	0,073	7,807	53	50,1	18,9	4,32	20	6,49	91,40
9	14/11/05	21,66	19,68	70	29	<0,1	3,68	1,13	4,81	0,307	0,164	0,182	10,5	14,23	72,5	21,8		55	7,06	21,54
10	15/12/05	27,71	20,14	210	32	0,1	5,35	0,66	6,01	0,311	0,369	0,974	11,03	15,64	116,1	21,6		30	7,13	12,29
11	14/03/06	27,69	6,96	120	29	0,1	6,79	4,68	11,47	0,046	0,916	0,223	7,7	18,55	114,3	23,4	3,22	20	6,803	
12	03/04/06	30,28	7,54	175	20	<0,1	5,77	1,02	6,79	0,024	1,103	0,888	9,78	14,96	73,4	21,1	1,8	40	7,03	
13	10/04/06	23,9	9,66	133	31	0,3	7,01	1,58	8,59	0,038	0,338	0,105	12,53	11,47	125,6	21,4	1,34	50	7,16	9,06
14	26/04/06	35,06	26,1	119	12	0,1	10,06	1,9	8,16	0,068	0,601	1,301	14,96	10,88	139,2	21,2	0,92	40	7,256	8,54
15	11/05/06	31,87	18,52	250	37	0,3	11,09	0,27	11,36	0,099		0,970	16,84	12,36	146,1	16,1	0,53	10	7,28	8,54
16	24/05/06	32,26	34,8	101	34	0,5	0,88	0,33	1,21	0,028	0,119	0,467	21,22	3,2	120	14,6	1,92	35	7,2	8,89
17	07/06/06	36,43	25,8	147	24	< 0,1	8,01	1,4	9,41	0,062	0,422	1,872	20,83	16,91	163,9	18	0,76	40	7	9,06
18	21/06/06	35,15	22,35	311	35	0,2	9,55	2,14	11,69	0,070	0,089	1,441	45,63	17,79	175,3	16,9	0,56	20	7,33	8,54
19	19/07/06	33,95	17,4	218	26	< 0,1	12,57	2,30	14,87	0,082		1,513	10,66	14,11	167,9	16,85	0,7	40	7,16	8,71
20	13/03/08	24,00	4,92	137	41	0,2	2,35	1,23	3,58	0,147	0,253	0,344	12,16	50,2	20,5	22	8,69	25	6,305	41,77
21	16/04/08	40,00	8,91	154	24	< 0,3	5,07	1,09	6,16	0,089		0,520	12,10	23,15	26,4	18,7	2,3	40	6,039	13,26
22	07/05/08	18,00	4,59	98	22	< 0,2	3,53	1,68	5,21	0,056	0,279	0,21	3,83	40,3	13,4	16,3	3,49	30	6,72	33,94
23	04/06/08	25,00	10,1	84	28	< 0,2	2,80	2,46	5,26	0,168	0,892	0,310	6,50	154	19	15,7	2,45	20	6,745	59,77
24	20/08/08	58,93	10,76	98	52	0,6	7,95	3,70	11,65	0,022		0,506	6,77	50,05	176	18,85	1,98	50	7,109	20,22

TABELA A5 - ESTAÇÃO DE MONITORAMENTO P4

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SS (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	17/06/05		51,6	16	28	<0,1	35,28	32,76	1,202	0,228	0,258	11,36	27,92	108,5	17,8	0,8	20	6,3	51,11
2	30/06/05	67,2	54	181	29	<0,1	55,44	15,96	1,334	0,118	0,254	11,5	21,77	136	18	0,1		6,69	19,62
3	20/07/05	40	54,0	249	40	<0,1	6,71	1,61	0,650	0,083	0,302	10,53	24,98	143,1	13,2	0,1		6,76	19,22
4	10/08/05	36,06	63,6	131	163	1,05	8,16	7,3696	0,445	0,827	0,290	6,75	124,00	69,2	12,9	5,4	10	6,72	100,03
5	14/09/05	27,17	12,9	160	28	<0,1	0,98	0,05	0,834	0,578	0,145	6,89	52	64,7	12,8		20	6,42	161,28
6	05/10/05	39,24	27,12	178	285	4,1	1,9	2,35	1,101	0,433	0,130	18,88	118	64,2	18,0		15	6,56	101,34
7	19/10/05	9,61	7,92	41	36	0,1	2,31	1,38	0,543	0,194	0,168	6,65	22,38	80,4	19,1	1,82	25	6,62	74,22
8	31/10/05	29,79	14,64		37	< 0,1			0,767	0,350	0,180	8,323		54,6	18,4	3,4	30	6,5	152,08
9	14/11/05	21,66	10,56	105,3	17	<0,1	4,05	1,22	0,564	0,204	0,298	5,801	14,02	92	22,3		45	7,13	30,16
10	15/12/05	26,17	17,7	230	22	0,1	4,88	0,28	0,406	0,361	0,608	10,55	13,3	125,5	21,6		40	7,16	25,01
11	14/03/06	43,07	9,84	125	71	1,7	8,4	2,26	0,074	0,878	0,905	12,35	58,46	123,5	22,7	2,12		7,399	22,57
12	03/04/06	12,75	11,2	167	13	<01	6,05	1,7	0,077	0,437	0,561	11,49	11,34	75,9	20,7	1,76	15	7,06	59,97
13	10/04/06	20,72	9,18	115	61	0,2	3,51	0,23	0,070	0,532	0,076	10,16	26,71	140,1	21,2	1,62		7,16	24,26
14	26/04/06	20,72	29,16	160	20	<0,1	8,1	1,63	0,050	2,639	1,266	13,53	26,28	157,1	21,3	1,34		7,295	17,03
15	11/05/06	43,03	23,53	247	36	0,4	9,82	0,27	0,104		1,114	18,28	16,44	157,3	16,2	0,37	25	7,28	17,69
16	24/05/06	33,87	18,2	335	70	1,5	0,88	0,71	0,520	0,144	0,388	16,95	3,4	122,1	14,4	1,28		7,2	20,97
17	07/06/06	31,45	8,53	129	51	0,2	7,56	1,29	0,051	0,948	1,553	21,37	16,87	166,9	18,2	0,66	30	7,1	17,03
18	21/06/06	33,70	12,4	416	12	< 0,1	8,45	1,81	0,048	0,082	1,417	16,32	14,32	181,2	17,1	0,5		7,38	14,73
19	19/07/06	61,08	25,2	199	102	1,2	10,15	1,76	0,027		0,007	36,47	74,00	168,4	16,87	1,32	15	7,135	16,37
20	13/03/08	26,00	5,94	150	45	0,5	2,46	1,18	0,170	0,210	0,326	18,53	48,5	21,6	21,1	8,2	25	6,394	103,31
21	16/04/08	36,80	3,8	164	30	< 0,2	5,18	1,20	0,135		0,397	10,23	30,5	29,4	18,5	1,76	30	6,192	30,16
22	07/05/08	22,00	6,02	122	22	< 0,2	4,26	1,79	0,069	0,297	0,23	3,40	44,35	15,8	15,9	2,56	35	6,51	108,57
23	04/06/2008	46,00	17,78	104	128	< 1,0	3,36	3,58	0,143	0,881	0,366	3,82	228	18	15,3	1,87	12	6,845	83,67
24	20/08/2008	24,38	7,19	102	24	0,4	8,29	3,14	0,049		0,387	6,14	40,5	197	18,9	1,05	45	7,147	21,63

TABELA A6 - ESTAÇÃO DE MONITORAMENTO P5

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SS (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	17/06/05		78	115	70	<0,1	28,73	19,32	0,493	0,137	0,546	8,819	22,82	93,7	17,6	1,1	35	6,20	69,40
2	30/06/05	40	49,2	167	22	<0,1	84,67	4,2	0,760	0,519	0,425	10,6	17,91	128,5	17,95	0,9		6,51	38,22
3	20/07/05	49,6	72,0	224	28	<0,1	28,68	6,14	0,595	0,344	0,281	8,63	15,51	129,2	13,3	1,2	60	6,70	26,70
4	10/08/05	36,06	39,3	159	212	1,05	9,48	9,48	0,380	0,693	0,344	7,77	143,00	73,5	13,1	3,4	15	6,71	71,80
5	14/09/05	9,06	9,34	112	25	<0,1	0,98	0,05	1,083	0,567	0,136	6,46	26,28	65,4	13,4		40	6,43	153,20
6	05/10/05	25,66	18,24	132	3	<0,1	1,85	1,06	0,886	0,167	0,171	8,53	66	65	18,3		15	6,38	97,10
7	19/10/05	18,6	9,84	72	41	0,3	2,02	1,1	0,641	0,234	0,141	7,26	25,72	73,4	19,7	2,16	50	6,59	84,08
8	31/10/05	17,25	13,2	92	29	< 0,1			0,736	0,304	0,057	7,029		55,7	19,1	3,24	70	6,52	166,60
9	14/11/05	24,99	9,36	180	9	<0,1	2,92	1,51	0,549	0,380	0,143	5,014	10,15	81,5	22,1		50	7,01	13,40
10	15/12/05	20,01	13,6	236	36	0,1	3,89	0,52	0,451	0,509	0,547	8,96	19,73	106,7	22		50	7,23	31,20
11	14/03/06	20	8,76	118	71	0,9	5,82	0,32	0,062	1,068	0,407	10,71	30,34	115,4	22,9	1,66		6,99	27,60
12	03/04/06	12,75	8,4	175	32	<0,1	4,64	1,41	0,068	0,308	0,370	11,14	21,38	69,4	20,6	2,14	40		76,20
13	10/04/06	17,53	7,02	186	41	<0,1	5,09	0,85	0,079	0,487	0,054	8,48	21,55	123,3	21,1	1,6	45	7,23	26,70
14	26/04/06	12,75	7,26	136	11	<0,1	4,78	1,58	0,064	0,550	0,820	9,25	16,19	129,9	20,9	2,14	60	7,24	19,30
15	24/05/06	29,03	6,45	188	47	0,5	0,6	0,05	0,053	0,152	0,232	10,77	13,77	108,1	14	1	35	7,20	27,00
16	07/06/06	22,95	8,83	94	35	< 0,1	6,61	1,18	0,045	0,198	1,226	13,3	12,19	150,8	18,3	0,98	60	7,00	19,98
17	21/06/06	22,79	13,29	155	15	<0,1	8,34	1,98	0,138	0,078	1,066	13,02	13,43	162,1	17,1	0,96	60	7,20	14,60
18	19/07/06	21,89	8,66	198	14	<0,1	9,27	0,93	0,043		0,804	32,37	14,49	150,3	16,45	2,28	60	7,06	15,80
19	13/03/08	26,00	3,10	129	48	< 0,5	2,02	1,18	0,149	0,299	0,320	18,20	52,5	17	23,5	6,53	35	6,46	192,92
20	16/04/08	38,40	9,09	142	44	< 0,5	3,95	1,15	0,085		0,338	10,81	34,05	24,4	18,7	1,26	100	6,73	15,20
21	07/05/08	18	3,07	102	24	< 0,1	3,86	1,01	0,089	0,395	0,148	3,28	39,55	14,3	16,2	2,41	30	6,59	131,86
22	04/06/08	54	17,43	72	168	< 1,3	6,61	2,91	0,108	0,676	0,457	8,82	221	19	15,4	1,37	15	6,92	56,80
23	20/08/08	43,69	6,03	120	28	< 0,4	6,83	2,69	0,050		0,338	5,43	41,7	156	19,1	1,51	50	6,99	81,00

TABELA A7 - ESTAÇÃO DE MONITORAMENTO P6

Coleta	Data	DQO (mg/L)	DBO_5 (mg/L)	SDT (mg/L)	SST (mg/L)	SS (mL/L)	N-A (mg/L)	N-Org (mg/L)	Nitrito (mg/L)	Nitrato (mg/L)	Fósforo (mg/L)	COT (mg/L)	Turbidez (NTU)	Condutividade (µS/cm)	T Água (°C)	OD (mg/L)	Secchi (cm)	pH	Vazão (m^3/s)
1	27/07/05	16	56,4	68	19	<0,1	5,53	4,39	0,334	1,204	0,132	6,53	29,87	64,4	13,6	3,4	30	6,30	128,70
2	10/08/05	23,52	37,0	76	66	<0,1	17,11	8,34	0,310	0,285	0,254	7,14	35,86	81	13,6	2,6	15	6,72	73,50
3	14/09/05	9,06	7,98	134	22	<0,1	0,66	0,6	0,831	0,487	0,046	4,88	28,36	48,3	13,2			6,22	175,04
4	05/10/05	18,11	18,72	155	104	0,1	0,73	0,11	0,890	0,323	0,035	8,24	86,5	46,3	18,3		10	6,32	120,20
5	19/10/05	9,34	7,44	81	28	<0,1	1,56	1,21	0,767	0,245	0,071	5,707	23,16	62,8	20	2,46	40	6,60	103,38
6	31/10/05	23,52	12,24	101	26	< 0,1			0,807	0,110	0,046	6,047		49,4	20,4	4,2	30	6,54	204,74
7	14/11/05	8,33	13,2	159	17	<0,1	2,64	1,04	0,901	0,500	0,128	3,846	5,03	76,2	23		60	7,16	27,27
8	15/12/05	16,93	6,82	246	37	0,1	3,18	0,52	0,965	0,525	0,337	5,82	13,6	93	23,1		50	7,14	35,19
9	14/03/06	18,46	7,92	149	43	0,4	4,69	2,26	0,093	0,650	0,265	9,14	23	101,7	23,4	2,66		6,99	30,78
10	03/04/06	14,34	10,42	148	33	<01	3,76	3,42	0,080	0,357	0,328	8,25	19,82	71,7	21		35	6,93	103,38
11	10/04/06	20,72	6,78	219	52	<0,1	5,94	0,06	0,099	0,426	0,020	6,9	10,71	109,5	22	2,96		7,08	29,73
12	26/04/06	12,54	6,9	135	16	<0,1	5,16	0,54	0,134	0,688	0,555	7,17	12,96	113,3	20,8	4,06	40	7,40	20,07
13	24/05/06	25,8	9,4	241	38	<0,1	6,09	0,55	0,028	0,616	0,306	22,65	6,97	99,9	14,7	1,16	40	7,30	30,44
14	07/06/06	16,85	9,6	53	21	< 0,1	5,66	1,01	0,058	0,080	0,958	9,12	12,83	135	17,8	2,6	50	7,10	8,11
15	21/06/06	17,65	5,54	255	5	< 0,1	10,15	1,04	0,093	0,069	0,909	10,4	11,00	149,9	17,6	2,56	60	7,47	16,47
16	19/07/06	19,04	5,74	178	15	< 0,1	8,23	1,21	0,055		0,634	30,35	14,45	135,2	17,13	1,28	45	7,24	11,78
17	13/03/08	28,00	3,14	93	25	< 0,5	1,57	1,19	0,129	0,505	0,341	11,75	80,0	14,2	24,1	7,79		5,17	77,62
18	16/04/08	30,40	6,26	146	14	< 0,2	3,11	0,92	0,123		0,255	7,18	28,25	21,7	18,7	1,08	90	6,12	53,74
19	07/05/08	13,00	3,00	128	16	< 0,1			0,125	0,529	0,121	3,30	30,3	12,7	16,4	3,06	40	6,71	171,51
20	04/06/08	21,00	< 2,00	106	8	< 0,3	5,49	0,22	0,063	0,416	0,387	8,96	81,5	20	15,1	2,1	15	7,06	56,36
21	20/08/08	38,61	3,05	122	20	< 0,2	5,6	2,24	0,068		0,251	8,88	24	137	19,45	2,16	45	7,06	113,04

APÊNDICE III – ESCORES DAS COMPONENTES PRINCIPAIS PARA ANÁLISE I

Ponto de Monitoramento	Amostra	ESCORES				
		CP1	CP2	CP3	CP4	CP5
PONTO 1	1	3,50	1,52	3,00	2,05	-0,89
	2	2,43	0,70	-0,91	0,23	0,78
	3	3,05	1,91	-0,94	0,74	0,75
	4	3,65	2,08	-0,03	1,10	0,44
	5	3,50	2,26	0,38	0,76	0,36
	6	4,17	2,89	0,36	0,96	-0,31
PONTO 2	7	2,09	-7,53	0,91	3,65	0,71
	8	-3,07	0,04	0,69	0,45	-0,58
	9	-2,02	0,94	-0,63	0,90	-0,15
	10	-2,37	0,12	0,35	0,72	-0,52
	11	-2,83	0,35	2,54	-0,40	1,35
PONTO 3	12	-1,06	-0,13	3,40	0,43	-3,75
	13	1,77	-1,10	1,43	-1,11	1,25
	14	-0,01	-2,20	-2,41	1,20	0,59
	15	-0,52	-0,03	-1,93	0,78	0,29
	16	-0,68	0,42	-0,56	-0,61	0,96
	17	-2,21	0,38	-1,24	0,66	-0,90
	18	-0,99	1,11	2,94	-1,18	1,81
	19	-2,58	0,09	0,08	0,71	-0,86
	20	-4,96	-0,58	0,95	-0,10	0,43
PONTO 4	21	2,30	-2,22	-0,31	-2,19	-0,91
	22	-0,22	-0,57	-1,81	-1,14	-0,18
	23	-2,46	-1,15	-0,47	0,57	0,97
PONTO 5	24	2,44	-2,31	0,63	-2,79	-0,55
	25	-0,18	-0,40	-1,42	-0,51	0,41
	26	-0,19	0,80	-1,94	-0,07	-0,17
	27	-0,36	0,18	2,09	-1,83	2,51
	28	-0,82	0,57	-0,35	-0,14	-0,47
	29	-1,38	0,78	-0,45	-0,15	-1,08
PONTO 6	30	3,02	-2,52	-0,54	-3,08	-1,74
	31	0,11	0,63	-2,01	-0,13	0,23
	32	-1,52	0,25	-0,13	-0,10	0,46
	33	0,04	1,12	-0,40	-0,01	-0,32
	34	-1,65	1,63	-1,29	-0,39	-0,93

APÊNDICE IV – QUADRADO DAS DISTÂNCIAS GENERALIZADAS E QUI-QUADRADOS RESPECTIVOS – ANÁLISE I

PONTOS PLOTADOS	d^2	χ^2
1	6,61	7,49
2	8,60	9,16
3	8,97	10,16
4	11,16	10,94
5	11,59	11,59
6	12,00	12,17
7	12,54	12,70
8	12,70	13,20
9	12,80	13,68
10	14,45	14,13
11	14,55	14,57
12	15,47	15,00
13	15,73	15,43
14	15,90	15,85
15	16,40	16,27
16	16,63	16,70
17	16,96	17,12
18	17,09	17,56
19	17,16	18,00
20	17,41	18,45
21	17,54	18,92
22	17,88	19,40
23	17,94	19,90
24	18,13	20,44
25	18,95	21,00
26	21,29	21,60
27	22,58	22,26
28	22,84	22,98
29	24,42	23,80
30	24,82	24,74
31	25,56	25,86
32	26,62	27,30
33	29,29	29,36
34	31,42	33,45

APÊNDICE V - ESCORES DOS NOVOS 5 PRIMEIROS FATORES – ANÁLISE I

Ponto de Monitoramento	Amostra	Data da Coleta	ESCORES				
			Fator 1	Fator 2	Fator 3	Fator 4	Fator 5
PONTO 1	1	20/07/2005	-2,23	-0,74	-0,13	0,44	1,31
	2	14/03/2006	-1,22	0,00	0,15	0,62	-0,37
	3	26/04/2006	-1,32	-0,59	-0,16	0,91	-0,69
	4	07/06/2006	-1,72	-0,69	-0,21	0,87	-0,42
	5	21/06/2006	-1,56	-0,66	-0,16	0,89	-0,44
	6	19/07/2006	-2,05	-1,03	-0,62	0,90	-0,13
PONTO 2	7	10/08/2005	-1,43	4,85	-0,63	0,17	0,53
	8	10/04/2006	1,19	0,09	-0,18	0,60	1,41
	9	24/05/2006	0,83	-0,64	-0,61	0,34	-0,21
	10	07/06/2006	0,72	0,32	-0,43	0,53	0,30
	11	21/06/2006	0,72	0,12	1,95	0,60	0,13
PONTO 3	12	20/07/2005	-0,08	-0,83	-0,64	-0,81	4,54
	13	19/10/2005	-0,91	0,51	1,78	-0,36	0,13
	14	14/03/2006	0,14	1,47	-0,91	0,26	-0,81
	15	03/04/2006	0,11	-0,21	-0,44	0,06	-0,62
	16	10/04/2006	0,47	-0,02	0,56	0,46	-0,34
	17	26/04/2006	0,98	-0,20	-1,02	0,30	0,38
	18	24/05/2006	0,22	-0,50	2,77	0,60	0,75
	19	07/06/2006	0,81	0,07	-0,50	0,35	0,73
	20	21/06/2006	1,71	0,92	0,64	0,94	0,59
PONTO 4	21	19/10/2005	-0,20	0,19	-0,08	-2,36	-0,35
	22	03/04/2006	0,58	-0,10	-0,48	-0,90	-0,96
	23	07/06/2006	0,96	0,64	0,59	0,30	-0,21
PONTO 5	24	19/10/2005	-0,51	-0,03	0,89	-2,63	-0,03
	25	10/04/2006	0,34	0,17	-0,18	-0,15	-0,68
	26	26/04/2006	0,30	-0,41	-0,77	0,11	-0,70
	27	24/05/2006	0,13	-0,16	2,79	-0,10	-0,85
	28	07/06/2006	0,50	-0,21	-0,45	0,05	-0,16
	29	21/06/2006	0,81	-0,31	-1,05	0,01	-0,08
PONTO 6	30	19/10/2005	-0,40	-0,30	-0,31	-3,47	-0,34
	31	26/04/2006	0,23	-0,43	-0,49	0,02	-0,86
	32	24/05/2006	0,79	-0,28	0,16	0,00	-0,49
	33	07/06/2006	0,21	-0,30	-0,55	0,34	-0,32
	34	21/06/2006	0,89	-0,72	-1,29	0,13	-0,72

ANEXOS

ANEXO I – FUNÇÃO PROGRAMADA NO MATLAB “COMP2”

```
function [ident,m,S,R,dd2,E2,CP,ESCR,RRYX] = comp(X)
close
close
close
close
close
%     ****************************************************************
%     *      ANÁLISE DE COMPONENTES PRINCIPAIS       *
%     ****************************************************************
%     *  Função programada pelo Prof. Jair Mendes Marques *
%     *         Departamento de Estatística da UFPR           *
%     ****************************************************************
%
%COMP  As componentes principais são combinações lineares das variáveis
%     originais, resultando num conjunto de variáveis não-correlaciona-
%     das, tendo propriedades especiais em termos de variâncias.
%        Os objetivos principais da ACP são:
%     (1) redução do número de variáveis;
%     (2) analisar quais as variáveis ou, quais os conjuntos de variá-
%        veis explicam a maior parte da variabilidade total revelando
%        que tipo de relacionamento existe entre as variáveis.
%
%     comp(X)  resulta em uma ACP das variáveis originais padronizadas,
%           ou seja, os autovalores e autovetores são obtidos da ma-
%           triz  de  correlações.  A matriz X é uma matriz de dados
%           (linhas = ítens,  colunas = variáveis).
%     [a,b,c,d,e,f,g,h,i] = comp(X) resulta na ACP como no caso anteri-
%           or, apenas que os argumentos de saída:  a=identificação,
%           b=vetor de médias, c=matriz covariância, d=matriz corre-
%           lação, e=autovalores, f=autovetores, g=comp. principais,
%           h=escores, i=correlações entre as CP e variáveis  origi-
%           nais, podem ser salvos para uso posterior.

% IDENTIFICAÇÃO
 ident='FUNÇÃO COMP/UFPR/DEPTO. DE ESTATÍSTICA/JMM';
% MEDIA-COV-CORRELACAO
  disp('    **********************')
  disp('    *   VETOR DE MÉDIAS   *')
  disp('    **********************')
  disp(' ')
  m=mean(X);
```

```
disp(m)
pause
disp('    *************************')
disp('    *   MATRIZ COVARIÂNCIA   *')
disp('    *************************')
disp('    ')
S=cov(X);
n1=length(diag(S));
if n1 < 8
   disp(S)
   pause
elseif n1 < 15
   disp(S(:,1:7))
   pause
   disp(S(:,8:n1))
   pause
elseif n1 < 22
   disp(S(:,1:7))
   pause
   disp(S(:,8:14))
   pause
   disp(S(:,15:n1))
   pause
 else
   disp(S)
pause
end
disp('    *************************')
disp('    *   MATRIZ CORRELAÇÃO   *')
disp('    *************************')
disp('   ')
R=corrcoef(X);
n2=length(diag(R));
if n2<8
   disp(R)
   pause
elseif n2 < 15
   disp(R(:,1:7))
   pause
   disp(R(:,8:n2))
   pause
elseif n2 < 22
   disp(R(:,1:7))
```

```
    pause
    disp(R(:,8:14))
    pause
    disp(R(:,15:n2))
    pause
  else
    disp(R)
    pause
  end
% AUTOVALORES E AUTOVETORES
  disp('    ************************************')
  disp('    * AUTOVALORES DA MATRIZ CORRELAÇÃO *')
  disp('    ************************************')
  disp(' ')
  [E2,D2]=eig(R);
  [dd2,i2]=sort(diag(D2));
  dd2=flipud(dd2)';
  i2=flipud(i2)';
  disp(dd2)
  pause
  disp('    ************************************');
  disp('    * AUTOVETORES DA MATRIZ CORRELAÇÃO *');
  disp('    ************************************');
  disp(' ')
  E2=E2(:,i2);
  [m2,n2]=size(E2);
  if n2 < 8
    disp(E2)
    pause
  elseif n2 < 15
    disp(E2(:,1:7))
    pause
    disp(E2(:,8:n2))
    pause
  elseif n2 < 22
    disp(E2(:,1:7))
    pause
    disp(E2(:,8:14))
    pause
    disp(E2(:,15:n2))
    pause
  else
    disp(E2)
```

```
    pause
  end
% COMPONENTES PRINCIPAIS DAS VAR. ORIGINAIS
  r1=eig(R);
  r1=flipud(sort(r1));
  m1=length(r1);
  j1=(1:m1)';
  t1=sum(r1);
  r2=(r1/t1)*100;
  r3=(cumsum(r1)/t1)*100;
  r=[j1 r1 r2 r3];
  disp('    ******************************************')
  disp('    * PROPORÇÃO DE VARIÂNCIA EXPLICADA PELOS *')
  disp('    *    AUTOVALORES DA MATRIZ CORRELAÇÃO    *')
  disp('    ******************************************')
  disp(' ')
  disp('   ----------------------------------------')
  disp('     ORDEM  AUTOVA-  VAR. EXPL.  VAR. EXPL. ')
  disp('             LORES     (EM %)    ACUM. (%)  ')
  disp('   ----------------------------------------')
  disp(sprintf('%8.0f %10.4f %8.2f %11.2f\n',r'))
  disp('   ----------------------------------------')
  pause
  disp('    ************************************************')
  disp('    * COMPONENTES PRINCIPAIS (VARIÁVEIS PADRONIZADAS) *')
  disp('    ************************************************')
  disp(' ')
  [E2,D2]=eig(R);
  [dd2,i2]=sort(diag(D2));
  dd2=flipud(dd2)';
  i2=flipud(i2)';
  E2=E2(:,i2);
  n2=length(dd2);
  if n2==1
    disp('   --------')
    disp('     CP1   ')
    disp('   --------')
    disp(E2)
    disp('   --------')
    pause
  elseif n2==2
    disp('   ------------------')
    disp('     CP1       CP2   ')
```

```
    disp('   ------------------')
    disp(E2)
    pause
elseif n2==3
    disp('   ----------------------------')
    disp('     CP1      CP2      CP3   ')
    disp('   ----------------------------')
    disp(E2)
    disp('   ----------------------------')
    pause
elseif n2==4
    disp('   ---------------------------------------')
    disp('     CP1      CP2      CP3      CP4   ')
    disp('   ---------------------------------------')
    disp(E2)
    disp('   ---------------------------------------')
    pause
elseif n2==5
    disp('   -------------------------------------------------')
    disp('     CP1      CP2      CP3      CP4      CP5   ')
    disp('   -------------------------------------------------')
    disp(E2)
    disp('   -------------------------------------------------')
    pause
elseif n2==6
    disp('   -----------------------------------------------------------')
    disp('     CP1      CP2      CP3      CP4      CP5      CP6   ')
    disp('   -----------------------------------------------------------')
    disp(E2)
    disp('   -----------------------------------------------------------')
    pause
elseif n2==7
    disp('   -------------------------------------------------------------------')
    disp('     CP1      CP2      CP3      CP4      CP5      CP6      CP7   ')
    disp('   -------------------------------------------------------------------')
    disp(E2)
    pause
elseif n2>7
    disp('   -------------------------------------------------------------------')
    disp('     CP1      CP2      CP3      CP4      CP5      CP6      CP7  ')
    disp('   -------------------------------------------------------------------')
    disp(E2(:,1:7))
    disp('   -------------------------------------------------------------------')
```

```
    pause
end
if n2==8
    disp('   --------')
    disp('     CP8   ')
    disp('   --------')
    disp(E2(:,8:n2))
    disp('   --------')
    pause
elseif n2==9
    disp('   ------------------')
    disp('     CP8       CP9   ')
    disp('   ------------------')
    disp(E2(:,8:n2))
    disp('   ------------------')
    pause
elseif n2==10
    disp('   ----------------------------')
    disp('     CP8       CP9      CP10   ')
    disp('   ----------------------------')
    disp(E2(:,8:n2))
    disp('   ----------------------------')
    pause
elseif n2==11
    disp('   ---------------------------------------')
    disp('     CP8       CP9      CP10      CP11   ')
    disp('   ---------------------------------------')
    disp(E2(:,8:n2))
    disp('   ---------------------------------------')
    pause
elseif n2==12
    disp('   -------------------------------------------------')
    disp('     CP8       CP9      CP10      CP11      CP12   ')
    disp('   -------------------------------------------------')
    disp(E2(:,8:n2))
    disp('   -------------------------------------------------')
    pause
elseif n2==13
    disp('   ------------------------------------------------------------')
    disp('     CP8       CP9      CP10      CP11      CP12      CP13   ')
    disp('   ------------------------------------------------------------')
    disp(E2(:,8:n2))
    disp('   ------------------------------------------------------------')
```

```
   pause
elseif n2==14
   disp('   ------------------------------------------------------------------')
   disp('    CP8      CP9     CP10     CP11     CP12     CP13     CP14   ')
   disp('   ------------------------------------------------------------------')
   disp(E2(:,8:n2))
   disp('   ------------------------------------------------------------------')
   pause
elseif n2>14
   disp('   ------------------------------------------------------------------')
   disp('    CP8      CP9     CP10     CP11     CP12     CP13     CP14   ')
   disp('   ------------------------------------------------------------------')
   disp(E2(:,8:14))
   disp('   ------------------------------------------------------------------')
   pause
end
if n2==15
   disp('   ---------')
   disp('    CP15   ')
   disp('   ---------')
   disp(E2(:,15:n2))
   disp('   ---------')
   pause
elseif n2==16
   disp('   ----------------------')
   disp('    CP15     CP16   ')
   disp('   ----------------------')
   disp(E2(:,15:n2))
   disp('   ----------------------')
   pause
elseif n2==17
   disp('   ------------------------------')
   disp('    CP15     CP16     CP17   ')
   disp('   ------------------------------')
   disp(E2(:,15:n2))
   disp('   ------------------------------')
   pause
elseif n2==18
   disp('   ---------------------------------------')
   disp('    CP15     CP16     CP17     CP18   ')
   disp('   ---------------------------------------')
   disp(E2(:,15:n2))
   disp('   ---------------------------------------')
```

```
    pause
  elseif n2==19
    disp('   ---------------------------------------------------')
    disp('     CP15      CP16      CP17      CP18      CP19   ')
    disp('   ---------------------------------------------------')
    disp(E2(:,15:n2))
    disp('   ---------------------------------------------------')
    pause
  elseif n2==20
    disp('   ------------------------------------------------------------')
    disp('     CP15      CP16      CP17      CP18      CP19      CP20   ')
    disp('   ------------------------------------------------------------')
    disp(E2(:,15:n2))
    disp('   ------------------------------------------------------------')
    pause
  elseif n2==21
    disp('   ---------------------------------------------------------------------')
    disp('     CP15      CP16      CP17      CP18      CP19      CP20      CP21   ')
    disp('   ---------------------------------------------------------------------')
    disp(E2(:,15:n2))
    disp('   ---------------------------------------------------------------------')
    pause
  elseif n2>21
    disp(E2)
    pause
  end
CP=E2;
    XM=mean(X);
    DP=diag(S);
    DP=(inv(sqrt(diag(DP))))';
    [m1,n1]=size(X);
    for i=1:m1
       AB(i,:)=X(i,:)-XM;
    end
    Z=AB*DP;
    ESCR=Z*E2;
    disp('    ************************************')
    disp('    * ESCORES (VARIÁVEIS PADRONIZADAS) *')
    disp('    ************************************')
    disp(' ')
    [m2,n2]=size(ESCR);
    if n2 < 8
      disp(ESCR)
```

```
      pause
    elseif n2 < 15
      disp(ESCR(:,1:7))
      pause
      disp(ESCR(:,8:n2))
      pause
    elseif n2 < 22
      disp(ESCR(:,1:7))
      pause
      disp(ESCR(:,8:14))
      pause
      disp(ESCR(:,15:n2))
      pause
    else
      disp(ESCR)
      pause
    end
  D2=diag(dd2);
  [m1,n1]=size(D2);
  RRYX=E2*sqrt(D2);
  disp('    ************************************************')
  disp('    * CORRELAÇÕES ENTRE AS VARIÁVEIS PADRONIZADAS *')
  disp('    *       E AS COMPONENTES PRINCIPAIS          *')
  disp('    ************************************************')
  disp(' ')
var=(1:n1)';
  RSYX1=[var RRYX];
  if n1<4
  disp('     ------------------------------')
  disp('     |   COMPONENTES PRINCIPAIS  |')
  disp('-----------------------------------')
    if n1==1
      disp('|VAR.|            CP1          |')
      disp('-----------------------------------')
      disp(sprintf('%3.0f %20.4f\n',RSYX1'))
      disp('-----------------------------------')
    elseif n1==2
      disp('|VAR.|       CP1     CP2        |')
      disp('-----------------------------------')
      disp(sprintf('%3.0f %13.4f %8.4f\n',RSYX1'))
      disp('-----------------------------------')
    elseif n1==3
      disp('|VAR.|    CP1      CP2      CP3  |')
```

```
        disp('-----------------------------------')
        disp(sprintf('%3.0f %10.4f %8.4f %8.4f\n',RSYX1'))
        disp('-----------------------------------')
      end
      pause
  elseif n1==4
  disp('    ----------------------------------------')
  disp('    |       COMPONENTES PRINCIPAIS      |')
  disp('--------------------------------------------')
  disp('|VAR.|   CP1     CP2     CP3     CP4  |')
  disp('--------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f\n',RSYX1'))
  disp('--------------------------------------------')
  pause
  elseif n1==5
  disp('    ------------------------------------------------')
  disp('    |          COMPONENTES PRINCIPAIS           |')
  disp('------------------------------------------------')
  disp('|VAR.|   CP1     CP2     CP3     CP4     CP5  |')
  disp('------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f\n',RSYX1'))
  disp('------------------------------------------------')
  pause
  elseif n1==6
  disp('    ----------------------------------------------------------')
  disp('    |              COMPONENTES PRINCIPAIS                |')
  disp('----------------------------------------------------------')
  disp('|VAR.|   CP1     CP2     CP3     CP4     CP5     CP6  |')
  disp('----------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',RSYX1'))
  disp('----------------------------------------------------------')
  pause
elseif n1==7
  disp('    ------------------------------------------------------------------')
  disp('    |                  COMPONENTES PRINCIPAIS                  |')
  disp('------------------------------------------------------------------')
  disp('|VAR.|   CP1     CP2     CP3     CP4     CP5     CP6     CP7  |')
  disp('------------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',RSYX1'))
  disp('------------------------------------------------------------------')
  pause
elseif n1==8
  disp('    ----------------------------------------------------------------------')
```

```
   disp('     |                    COMPONENTES PRINCIPAIS                    |')
   disp('-------------------------------------------------------------------------------')
   disp('|VAR.|    CP1      CP2      CP3      CP4      CP5      CP6      CP7      CP8  |')
   disp('-------------------------------------------------------------------------------')
   disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',RSYX1'))
   disp('-------------------------------------------------------------------------------')
   pause
elseif n1>8
   R1=(RSYX1(:,[1:9]))';
   R2=(RSYX1(:,[1 10:n1+1]))';
   disp('     ----------------------------------------------------------------------------------')
   disp('     |                         COMPONENTES PRINCIPAIS                          ')
   disp('----------------------------------------------------------------------------------')
   disp('|VAR.|    CP1      CP2      CP3      CP4      CP5      CP6      CP7      CP8  |')
   disp('----------------------------------------------------------------------------------')
   disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',R1))
   disp('----------------------------------------------------------------------------------')
   pause
end
   if n1==9
   disp('---------------')
   disp('|VAR.|    CP9  ')
   disp('---------------')
   disp(sprintf('%3.0f %10.4f\n',R2))
   disp('---------------')
   pause
   elseif n1==10
   disp('--------------------------')
   disp('|VAR.|    CP9     CP10  |')
   disp('--------------------------')
   disp(sprintf('%3.0f %10.4f %8.4f\n',R2))
   disp('--------------------------')
   pause
   elseif n1==11
   disp('-----------------------------------')
   disp('|VAR.|    CP9     CP10     CP11  |')
   disp('-----------------------------------')
   disp(sprintf('%3.0f %10.4f %8.4f %8.4f\n',R2))
   disp('-----------------------------------')
   pause
      elseif n1==12
   disp('-------------------------------------------')
   disp('|VAR.|    CP9     CP10     CP11     CP12  |')
```

```
  disp('-------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f\n',R2))
  disp('-------------------------------------------')
  pause
     elseif n1==13
  disp('-----------------------------------------------------')
  disp('|VAR.|    CP9     CP10     CP11     CP12     CP13  |')
  disp('-----------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f\n',R2))
  disp('-----------------------------------------------------')
  pause
     elseif n1==14
  disp('-------------------------------------------------------------')
  disp('|VAR.|    CP9     CP10     CP11     CP12     CP13     CP14  |')
  disp('-------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f%8.4f %8.4f\n',R2))
  disp('-------------------------------------------------------------')
  pause
     elseif n1==15
  disp('---------------------------------------------------------------------')
  disp('|VAR.|    CP9     CP10     CP11     CP12     CP13     CP14     CP15  |')
  disp('---------------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',R2))
  disp('---------------------------------------------------------------------')
  pause
     elseif n1==16
  disp('-----------------------------------------------------------------------------')
  disp('|VAR.|    CP9     CP10     CP11     CP12     CP13     CP14     CP15     CP16  |')
  disp('-----------------------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',R2))
  disp('-----------------------------------------------------------------------------')
  pause
     elseif n1>16
  R3=(RSYX1(:,[1 10:17]))';
  R4=(RSYX1(:,[1 18:n1+1]))';
  disp('-----------------------------------------------------------------------------')
  disp('|VAR.|    CP9     CP10     CP11     CP12     CP13     CP14     CP15     CP16  |')
  disp('-----------------------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.3f %8.4f\n',R3))
  disp('-----------------------------------------------------------------------------')
  pause
end
if n1==17
```

```
    disp('---------------')
    disp('|VAR.|    CP17  ')
    disp('---------------')
    disp(sprintf('%3.0f %10.4f\n',R4))
    disp('---------------')
    pause
  elseif n1==18
    disp('--------------------------')
    disp('|VAR.|    CP17     CP18  |')
    disp('--------------------------')
    disp(sprintf('%3.0f %10.4f %8.4f\n',R4))
    disp('--------------------------')
    pause
  elseif n1==19
    disp('-----------------------------------')
    disp('|VAR.|    CP17     CP18     CP19  |')
    disp('-----------------------------------')
    disp(sprintf('%3.0f %10.4f %8.4f %8.4f\n',R4))
    disp('-----------------------------------')
    pause
  elseif n1==20
    disp('-------------------------------------------')
    disp('|VAR.|    CP17     CP18     CP19     CP20  |')
    disp('-------------------------------------------')
    disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f\n',R4))
    disp('-------------------------------------------')
    pause
  elseif n1==21
    disp('---------------------------------------------------')
    disp('|VAR.|    CP17     CP18     CP19     CP20     CP21  |')
    disp('---------------------------------------------------')
    disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f\n',R4))
    disp('---------------------------------------------------')
    pause
  elseif n1==22
    disp('-----------------------------------------------------------')
    disp('|VAR.|    CP17     CP18     CP19     CP20     CP21     CP22  |')
    disp('-----------------------------------------------------------')
    disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',R4))
    disp('-----------------------------------------------------------')
    pause
  elseif n1==23
    disp('-------------------------------------------------------------------')
```

```
  disp('|VAR.|    CP17     CP18     CP19     CP20     CP21     CP22     CP23  |')
  disp('-------------------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',R4))
  disp('-------------------------------------------------------------------------')
  pause
elseif n1==24
  disp('-----------------------------------------------------------------------------------')
  disp('|VAR.|    CP17     CP18     CP19     CP20     CP21     CP22     CP23     CP24  |')
  disp('-----------------------------------------------------------------------------------')
  disp(sprintf('%3.0f %10.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f %8.4f\n',R4))
  disp('-----------------------------------------------------------------------------------')
  pause
end
    n2=length(dd2);
    x2=1:n2;
    figure(1)
    clf
    plot(x2,dd2,x2,dd2,'o')
    grid
    title('AUTOVALORES DA MATRIZ CORRELAÇÃO')
    xlabel('NÚMERO DO AUTOVALOR')
    ylabel('AUTOVALOR')
    pause
    n2=length(E2(:,1));
    figure(2)
    clf
    plot(E2(:,1),E2(:,2),'r.','markersize',15)
    grid
    for i=1:n2
       text(E2(i,1),E2(i,2)+0.05,num2str(i))
    end
    title('COMPONENTES PRINCIPAIS: COMPON 1 versus COMPON 2')
    xlabel('COMPONENTE 1')
    ylabel('COMPONENTE 2')
    pause
      figure(3)
      clf
      n2=length(ESCR(:,1));
      plot(ESCR(:,1),ESCR(:,2),'r.','markersize',15)
      grid
      for i=1:n2
         text(ESCR(i,1),ESCR(i,2)+0.1,num2str(i),'fontsize',10)
      end
```

```
title('DISPERSÃO DOS ESCORES: COMP1 versus COMP2')
xlabel('ESCORE - COMP1')
ylabel('ESCORE - COMP2')
```

ANEXO II – FUNÇÃO PROGRAMADA NO MATLAB “NORMULT”

```
function [d2,q2] = normult(x)
% Função destinada a verificar a normalidade multivariada da amostra x.
% d2 = distância quadrática
% q2 = qui-quadrado

[n,p] = size(x);
m = mean(x);
S = cov(x);

% Cálculo das Distâncias Generalizadas
for i=1:n
   d2(i)=(x(i,:)-m)*inv(S)*(x(i,:)-m)';
end
d2 = sort (d2); % coloca em ordem crescente

% Cálculo do q2 (qui-quadrado)
for i=1:n
   q2(i)=chi2inv((i-0.5)/n,p); % p é o grau de liberdade, i é o valor da área
end

% Diagrama
plot(d2,q2,'*b') % b é a primeira letra do nome da cor, nesse caso é blue
xlabel('d^2')
ylabel('\chi^2')
grid
```

ANEXO III – FUNÇÃO PROGRAMADA NO MATLAB “KMO”

```
function y = KMO(X)
% Função que tem o objetivo de calcular a Estatística
% de Bartlett para o teste de esfericidade e a Medida
% de Adequacidade da Amostra de Kaiser-Meyer-Olkin. O
% argumento de entrada é: X = matriz de dados(amostra
% multivariada).
R=corrcoef(X);
[n,p]=size(X);
% Cálculo da estatística de Bartlett
Q2=-((n-1)-(2*p+5)/6)*log(det(R));
GL=p*(p-1)/2;
pvalor=(1-chi2cdf(Q2,GL));
disp('Teste de Esfericidade -Estatística de Bartlett')
disp(' ')
Q2
disp(' ')
pvalor
disp(' ')
% Cálculo da medida KMO
[p,p]=size(R);
for i=1:p-1
      for j=i+1:p
       l=0;
        for k=1:p
          if (i~=k)&(j~=k)
            l=l+1;
            w(l)=k;
          else
            m=1;
          end
        end
    Y1=X(:,i);
    X1=X(:,w);
    B1=pinv(X1'*X1)*(X1'*Y1);
    e1=Y1-X1*B1;
    Y2=X(:,j);
    B2=pinv(X1'*X1)*(X1'*Y2);
    e2=Y2-X1*B2;
    r(i,j)=sum(e1.*e2)/sqrt((sum(e2.^2))*(sum(e1.^2)));
    r(j,i)=r(i,j);
```

```
      r(i,i)=0;
      clear w
       end
end
q=r;
r2=R.^2;
q2=q.^2;
sr2=0;
sq2=0;
for i=1:p
   for j=1:p
      if i==j
         k=1;
      else
         sr2=sr2+r2(i,j);
         sq2=sq2+q2(i,j);
      end
   end
end
MSA=sr2/(sr2+sq2);
disp('Medida de adequacidade da amostra de Kaiser-Meyer-Olkin')
disp(' ')
MSA
```

ANEXO IV – FUNÇÃO PROGRAMADA NO MATLAB “COPHENET”

```
function c = cophenet(Z,Y)
%COPHENET Cophenetic coefficient.
%   C = COPHENETIC(Z,Y) computes the Cophenetic coefficient between the
%   distance of the cluster tree in Z and the distance in Y. Z is the
%   output of the function LINKAGE. Y is the output of the function
%   PDIST.
%
%   The Cophenetic coefficient is defined as
%
%                 sum((Z(i,j)-z)*(Y(i,j)-y))
%                 i<j
%       c =   ----------------------------------------
%           sqrt(sum((Z(i,j)-z)^2)*sum((Y(i,j)-y)^2))
%                i<j              i<j
%
%   Y(i,j) is the distance between observation i and j. y is mean(Y).
%   Z(i,j) is the distance between observation i and j at the combine
%   time and z = mean(Z).
%
%   See also PDIST, LINKAGE, INCONSISTENT, DENDROGRAM, CLUSTER, CLUSTERDATA

%   ZP You, 3-10-98
%   Copyright (c) 1993-98 by The MathWorks, Inc.
%   $Revision: 1.2 $

n = size(Z,1)+1;

link = zeros(n,1); listhead = 1:n;
sum1 = 0; sum2 = 0; s11 = 0; s22 = 0; s12 = 0;

for k = 1:(n-1)
  i = Z(k,1); j = Z(k,2); t = Z(k,3);
  m1 = listhead(i); % head of the updated cluster i
  while m1 > 0
    m = listhead(j);
    while m > 0
      u = Y((m1-1)*(n-m1/2)+m-m1); % distance between m and m1.
      sum1 = sum1+t; sum2 = sum2+u;
      s11 = s11+t*t; s22 = s22+u*u;
      s12 = s12+t*u;
      msav = m;
      m = link(m);
    end
    m1 = link(m1); % find the next point in cluster i
  end

  % link the end of cluster j to the head of cluster i
  link(msav) = listhead(i);

  % make the head of newly formed cluster i to be the head of cluster
  % j before the merge.
  listhead(n+k) = listhead(j);

end
t = 2/(n*(n-1));
s11 = s11-sum1*sum1*t; s22 = s22-sum2*sum2*t; s12 = s12-sum1*sum2*t;
c = s12/sqrt(s11*s22); % cophenectic coefficient formula
```